Working Paper Series

Averaging Impulse Responses Using Prediction Pools

WP 23-04

Paul Ho
Federal Reserve Bank of Richmond

Thomas A. Lubik
Federal Reserve Bank of Richmond

Christian Matthes
Indiana University

FEDERAL RESERVE BANK OF RICHMOND®
Richmond ▪ Baltimore ▪ Charlotte

Averaging Impulse Responses Using Prediction Pools*

Paul Ho
Federal Reserve Bank of Richmond†

Thomas A. Lubik
Federal Reserve Bank of Richmond‡

Christian Matthes
Indiana University§

January 25, 2023

Abstract

Macroeconomists construct impulse responses using many competing time series models and different statistical paradigms (Bayesian or frequentist). We adapt optimal linear prediction pools to efficiently combine impulse response estimators for the effects of the same economic shock from this vast class of possible models. We thus alleviate the need to choose one specific model, obtaining weights that are typically positive for more than one model. Three Monte Carlo simulations and two monetary shock empirical applications illustrate how the weights leverage the strengths of each model by (i) trading off properties of each model depending on variable, horizon, and application and (ii) accounting for the full predictive distribution rather than being restricted to specific moments.

JEL Classification: C32, C52

Key Words: Prediction Pools, Model Averaging, Impulse Responses, Misspecification

*We are grateful to Mikkel Plagborg-Møller, Mark Watson, our discussant Ed Herbst, and workshop participants at the 2021 CEF conference, the Drautzburg-Nason workshop, the Advances in Macro-Econometrics Conference, the 2022 System Econometrics Meeting, the 2022 Midwest Econometrics meeting, Indiana University and the Federal Reserve Bank of Richmond for helpful comments. Aubrey George, Colton Lapp, and Brennan Merone provided excellent research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

†Research Department, P.O. Box 27622, Richmond, VA 23261. Email: paul.ho@rich.frb.org.

‡Research Department, P.O. Box 27622, Richmond, VA 23261. Email: thomas.lubik@rich.frb.org.

§Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405. Email: matthesc@iu.edu.

1 Introduction

Impulse responses are a key tool in macroeconomists' arsenal to trace out the effects of structural shocks on aggregate quantities and prices. When estimating these impulse responses, economists have a wide range of options. For example, in the space of purely statistical models, a researcher can choose, between local projections (LPs) and vector autoregressions (VARs), Bayesian and frequentist methods, and different specifications. Naturally, each choice has its own drawbacks and benefits. It is well known that these choices can generate significantly different results (see Ramey (2016) for several leading examples). While there is a growing literature discussing conditions under which one approach might be preferred over another (Stock and Watson, 2018; Herbst and Johannsen, 2020; Plagborg-Møller and Wolf, 2021), in practical applications many of the conditions are likely difficult to verify.

In this paper, we introduce a method to average impulse responses from different estimators by extending the optimal prediction pools studied by Geweke and Amisano (2011) and Hall and Mitchell (2007).[1] In particular, we compute the optimal weights that maximize the weighted average log score function for forecasts conditional on the structural shock of interest. The only input required is a set of forecast densities that trace out the model-specific effects of the shock of interest. The individual impulse responses can be based upon any method that delivers such a conditional forecast density for a given variable at a given horizon.

Our approach is designed to appeal to empirical macroeconomists, who may find it difficult to choose between different methods for estimating impulse responses. LPs have become popular because they allow for introduction of extraneous variables in a straightforward manner. At the same time, confidence intervals of LP-based responses tend to be wide and cannot strictly be interpreted as structural. In contrast, VAR-based impulse response estimates suffer from a well-known bias. Our proposed solution to these issues in practical applications is to simply take all of these concerns at face value and compute combined responses that take these trade-offs into account. Our use of prediction pools provides a systematic and computationally tractable method to account for these issues in a wide range of applications, where the focus is on identifying plausible and robust dynamic behavior over time, irrespective of the underlying models.

A key strength of our approach is its flexibility. In particular, it removes the necessity to choose one model or even one statistical paradigm. Moreover, the methodology is applicable to a wide range of models. Typical methods such as Bayesian model averaging are

[1] Opinion pools, i.e., a forecast density formed by averaging over model-specific forecast densities, were first introduced by Stone (1961).

unavailable when one of the estimators considered is based on LPs as LPs are not 'generative models', that is, a set of LPs for different horizons do not form a consistent data-generating process. Besides the aforementioned LPs and VARs, dynamic equilibrium models (Smets and Wouters, 2007), dynamic factor models (Stock and Watson, 2016), or single equation methods (Baek and Lee, 2022) can be used in our framework. In addition, our method provides horizon- and variable-specific averages, thus exploiting each method's strength as much as possible.

As highlighted by Geweke and Amisano (2011), prediction pools have properties that make them well-suited to average over models or estimators when it is clear that all included models are misspecified. In contrast to Bayesian model averaging or related frequentist methods, more than one model will generally receive a positive weight. This helps prediction pools to outperform other model selection or model averaging approaches using various measures of forecast accuracy. Our extension inherits these properties, as we demonstrate with Monte Carlo exercises and with two applications that study the effects of U.S. monetary policy shocks.

Prediction pools are also computationally straightforward to implement relative to alternative methods of averaging across models. Since each model-specific forecasting density can be obtained separately, the most-time consuming part of forming prediction pools can be parallelized. The second step is then a relatively simple numerical maximization problem with a concave objective function and convex constraints. Other methods that also combine information from various models such as mixture models or composite-likelihood estimators (Qu, 2018; Canova and Matthes, 2021) do not share this modularity and thus have substantially higher computational complexity.

Overall, our paper highlights several broad messages for estimating impulse responses. The theoretical properties of individual models are not sufficient criteria for the choice of optimal weights in the prediction pool. Misspecified models can dominate correctly specified (or more flexible) models in finite samples. On the other hand, models that produce tighter estimates need not receive greater weight. The choice of model and their weights depend on the entire predictive distribution and not only the point estimates. While our examples focus on the mean and variance, higher-order moments or any other properties of the predictive distribution can be important more generally. Finally, we also find that the optimal weights on models depend on horizon, variable, and application, making it difficult to derive general guidelines or rules-of-thumb.

We illustrate our methodology using three Monte Carlo experiments. The first is a stylized univariate example motivated by Herbst and Johannsen (2020), which serves as a proof of concept and implies reasonable optimal weights. They result in an average impulse

response with a bias similar to one chosen by minimizing a squared error criterion. The second Monte Carlo compares a VAR and an LP in a setting where the VAR is misspecified, but the LP produces noisier estimates and has finite sample bias that is of the opposite sign from the VAR. While most of the weight is placed on the VAR, substantial weight is also placed on the LP, reducing the bias of the averaged impulse response relative to the VAR on its own, with the biases of the two models offsetting each other. The final Monte Carlo simulates data from the DSGE model from Smets and Wouters (2007), illustrating the weighting scheme's ability to trade off bias and variance for a relatively realistic data-generating process. These exercises highlight how our approach often gives positive weight to all competing models, but is also consistent with previous theoretical results found in Herbst and Johannsen (2020) and Li et al. (2022).

We then consider two empirical exercises. Our first application uses an instrument that exploits high-frequency variation in asset prices around monetary policy decisions (Gertler and Karadi, 2015; Caldara and Herbst, 2019). The second application follows Ramey (2016), where we average across four models that use the same Romer and Romer (2004) narrative instrument for monetary shocks. We find a range of results depending on application, horizon, and variable, emphasizing the flexibility of our methodology and the importance of considering the full predictive distribution rather than individual statistics. In addition, we find cases where the averaged impulse response delivers an economic message that is different and often more plausible than any of the individual models.

Related Literature. Our approach is motivated by the vast array of choices for computing impulse responses available to practitioners. It is designed to allow researchers to average optimally across multiple approaches rather than choosing just one. The two main statistical models are VARs (Sims, 1980) and LPs (Jordà, 2005), which we focus on in our Monte Carlos and applications. Within these two classes of models there are numerous variations. For example, in VARs inference can be conducted using Bayesian or frequentist methods (Sims and Zha, 1999). The Bayesian approach requires the choice of priors (Doan et al., 1984; Del Negro and Schorfheide, 2004; Giannone et al., 2015) while the frequentist approach requires choices about bias correction and the construction of confidence intervals (Kilian, 1998; Pesavento and Rossi, 2006). With LPs, there is a growing literature providing choices on the approach to inference (Herbst and Johannsen, 2020; Montiel Olea and Plagborg-Møller, 2021; Lusompa, 2021; Bruns and Lütkepohl, 2022) and smoothing the impulse responses (Barnichon and Brownlees, 2019; Miranda-Agrippino and Ricco, 2021a).

Having a general method that is flexible enough to cater to different models, variables, and horizons is particularly useful given the range of conclusions in the literature about the

relative strengths of the different methods. While there are asymptotic results on the relative performance of VARs and LPs (Stock and Watson, 2018; Plagborg-Møller and Wolf, 2021), the conditions for these theorems may not be easily verifiable in practice. In finite sample settings, the literature has also compared the performance of VARs and LPs (Kilian, 1998; Marcellino et al., 2006; Li et al., 2022). However, it is difficult to draw general conclusions, especially in empirical applications when the true model is not known. Our Monte Carlos and empirical applications show that the relative weights on different models can vary drastically not only with the data but also by variable and horizon. We consider it therefore important to rely on a general method that is able to assign weights variable by variable and horizon by horizon.

Prediction pools have been used to average models since their introduction by Geweke and Amisano (2011) and subsequent follow-up work in Geweke and Amisano (2012) and Amisano and Geweke (2017). The methodology has been extended by Waggoner and Zha (2012) and Del Negro et al. (2016) to assign time-varying weights. Our key innovation is that prediction pools can be used to average impulse responses whereby we treat the impulse responses as conditional forecasts. This allows for a flexible method that inherits desirable properties of the original prediction pools.

Model averaging has a long tradition in economics, partially motivated by the observation that averages of forecasts across multiple models tend to outperform forecasts based on an individual model (Bates and Granger, 1969). This is, in fact, one of the original motivations behind optimal prediction pools (Geweke and Amisano, 2011). Alternative averaging methods exist in both Bayesian and frequentist frameworks. In the Bayesian setting, model averaging is just another application of Bayes' theorem (for an application to VARs, see, for example, Strachan and van Dijk (2007)). As mentioned before, Bayesian model averaging generally requires use of generative models and, as such, rules out the use of LPs. Frequentist versions of model or forecast averaging such as Hansen (2007) also focus on specific classes of models (averages of least squares estimators in that case). Hansen (2016) studies model combination of various restricted VARs estimated via least squares. He proposes to find optimal model weights to minimize the mean squared error of a function of the VAR parameters (which can be an impulse response at a specific horizon).

Outline. The rest or the paper is structured as follows. Section 2 introduces our methodology. Section 3 describes our three Monte Carlo exercises. In Section 4, we apply our method to study the response of various macroeconomic aggregates to monetary shocks identified by the instruments from Gertler and Karadi (2015) and Romer and Romer (2004). Section 5 concludes.

2 Prediction Pools

We use prediction pools to average impulse responses across different models, based on Geweke and Amisano (2011). In their framework, predictive densities $p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)$ for each model $\mathcal{M}_m$ are combined to create a predictive density for an observable z_t conditional on model-specific predictive variables X_m^t, from which objects of interest, such as forecasts, can be computed.[2] The individual predictive densities are taken as given; that is, in contrast with other approaches to model averaging such as the estimation of mixture models, the parameters of the specific models and the model weights are not estimated jointly.

Formally, for any given horizon h, the goal is to maximize the log predictive score function:

$$\max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_{m,h} p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)\right], \tag{1}$$

where z_{t+h} denotes the variable of interest, X_m^t denotes the history of variables that z_{t+h} depends on in model $\mathcal{M}_m$, and $m = 1, ..., M$ indexes different models. The framework can be extended to the multivariate case, where z_{t+h} can be a vector of observables, but for ease of exposition and in our empirical setting later on we find it useful to focus on one variable at a time.[3]

2.1 Adapting Prediction Pools to Impulse Response Averaging

Prediction pools generally improve forecasting ability relative to individual models as judged by the log predictive score (Geweke and Amisano, 2011, 2012). They do so by usually giving more than one model a positive weight, in contrast with posterior model probabilities in a Bayesian setting.

We leverage the useful properties of prediction pools for the problem of impulse response estimation using the insight that impulse responses are nothing but conditional forecasts.

[2] Subscripts generally denote period-specific outcomes (except for the subscript m, which denotes the model at hand), whereas superscripts denote histories up to and including the period specified in the superscript.

[3] We treat the computation of the weights at different horizons as distinct problems. One could, alternatively, compute weights jointly and impose a penalty that forces the changes in the weights across horizons to be smoother than in our benchmark. For example, one could estimate the sets of weights $\{\{w_{m,h}\}_{m=1}^{M}\}_{h=1}^{H}$ by maximizing the following objective function:

$$\max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{h=1}^{H}\sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_{m,h} p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)\right] + \lambda \sum_{h=2}^{H}\sum_{m=1}^{M} (w_{m,h} - w_{m,h-1})^2,$$

where λ controls how much smoother the weights will be relative to our benchmark. With $\lambda = 0$, we replicate our benchmark since then each horizon's weights can be solved for independently of all other horizons.

The impulse responses in our framework are averages of model-specific impulse response estimators.[4] Our approach thus rewards models that forecast well.

The primitives that we need for our approach are forecasting densities based on each model. Where we differ from Geweke and Amisano (2011) is that we use a measure of the shock of interest as a conditioning argument in our predictive densities. In general, we form forecast densities that depend on observables up to time $t-1$ and a measure of the structural shock at time t. These measures of shocks can depend on time t data and model parameters. They can also incorporate identification restrictions, as will become clear in our examples. Whereas the forecasting densities used in Geweke and Amisano (2011) and other papers that use prediction pools make no explicit use of identified shocks or specific identification schemes for these shocks, incorporating these in our approach allows us to discriminate between models with different identification schemes.[5]

This conditioning scheme makes our approach more relevant for the empirical practice and choices that researchers are facing. An alternative approach would be to use the Geweke and Amisano (2011) approach for different reduced-form models directly and then impose identifying restrictions ex-post after finding the optimal weights. However, many applications of LPs directly use information on structural shocks (or instruments thereof) in the estimation, making this alternative less appealing when at least one of the models in our pool is based on LPs.

Another distinctive component of our approach is how we implement the distribution over each model's parameters, i.e., how our forecasting densities incorporate parameter uncertainty within a model. Geweke and Amisano (2011) use two approaches: Either the posterior distribution of parameters from a Bayesian estimation or, alternatively, fixed parameter values from some point estimate. We use a more general framework where the parameters of model $\mathcal{M}_m$ are collected in a vector Ω_m. We generate draws from a distribution $p_m(\Omega_m)$ that captures the parameter uncertainty we want to consider. This could be a posterior distribution, a point mass, a prior distribution, or a distribution derived using frequentist principles, say by appealing to standard asymptotic arguments or numerical approaches such as the bootstrap.

We generally study a vector y_t of macroeconomic variables and denote the jth variable of that vector by $y_{t,j}$. Since LPs are usually estimated for one specific variable and horizon at a time, we carry out our analysis variable by variable and horizon by horizon as well. This also gives us additional flexibility as different models might yield better forecasting ability

[4] Our focus on this paper is on linear models, but our approach could also be used in nonlinear settings.

[5] For the different identification schemes to be comparable, we are implicitly assuming that they are indeed identifying the same shock.

for different variables or horizons.

With these definitions in hand, we define our forecasting density for model m, p_m^*:

$$p_m^*\left(y_{t+h,j}\right) = \int p\left(y_{t+h,j}|y^{t-1}, \varepsilon_t(\Omega_m, y^t), \Omega_m, \mathcal{M}_m\right) p_m(\Omega_m) d\Omega_m, \tag{2}$$

which replaces the forecasting densities $p\left(z_{t+h}|X_m^t; \mathcal{M}_m\right)$ in Equation (1). We can approximate the integral on the right-hand side by Monte Carlo methods, as is often necessary in practice. We can extend our definition of p^* by allowing different models to depend on different right-hand side variables. While we allow the shock measure $\varepsilon_t(\Omega, y^t)$ to be model-specific, in our applications we use the same shock (or instrument of a shock as a conditioning argument) in all models and assume that the observed shock is one element of the vector y_t.

Geweke and Amisano (2011) use true out-of-sample forecasting densities, i.e., their densities $p\left(z_{t+h}|X_m^t; \mathcal{M}_m\right)$ are generally re-estimated every period. While this is possible in our framework as well, we use an alternative approach inspired by cross validation. In particular, we split the sample in half and estimate the models for each subsample separately. We then use the implied out-of-sample forecasting densities for the parts of the sample that were not used for estimation to obtain model weights. More specifically, we first estimate each model using the first half of the sample, and then use those parameter estimates to forecast the second half. In the next step, we estimate using the second sub-sample, fix parameter estimates and forecast the first subsample. This produces two true out-of-sample forecast densities without having to re-estimate every period. We view this approach as trading off computing time and overfitting concerns, which would play a role if we did not split the sample at all.[6]

2.2 Properties of Prediction Pools

With forecasting densities (2) in hand, the theorems stated in Geweke and Amisano (2011) all apply. In particular, as long as the expected average forecast densities do not take on the same value for different models, the true model, should it be contained in the set of models we consider, asymptotically receives a weight of 1. In contrast to Bayesian model averaging, more than one model will receive positive weight even asymptotically if the true model is not contained in the set of weights (Geweke and Amisano, 2012). The individual model with the highest log predictive score might not even receive a positive weight in the

[6] We could extend our approach to allow for time-varying weights along the lines of Waggoner and Zha (2012) or Del Negro et al. (2016).

optimal pool if there are more than two models being considered.[7] Furthermore, the weights satisfy a number of consistency requirements that make their use appealing. We state these consistency requirements as derived by Geweke and Amisano (2011) in Appendix B.

We consider the prediction pool framework as particularly well suited for applications in empirical macroeconomics. First, by studying each horizon separately, we overcome the issue that LPs are not *generative* models. In particular, there is no unique way to simulate a sample of arbitrary length from LPs estimated using different horizons. The simulation from one horizon is in general inconsistent with simulations from LPs for a different horizon. As a result, Bayesian model averaging is not possible. Second, prediction pools allow us to compare Bayesian and frequentist approaches. In particular, the probability distribution $p_m(\Omega_m)$ can be either Bayesian (i.e., a posterior distribution) or frequentist (i.e., an asymptotic distribution).[8] Finally, solving the optimization problem (1) is computationally straightforward.

2.3 Implementation

We now present a step-by-step guide that summarizes our approach.

1. Split the estimation sample in half, so that each subsample has $T/2$ observations (we assume for simplicity that T is even). We denote the subsample by $s = 1, 2$, where $s = 1$ means that periods dated $t = 1, ...T/2$ are used in the estimation, whereas $s = 2$ means that periods $t = T/2 + 1, ..., T$ are used. In a slight abuse of notation, we define a function $s(t)$ that is equal to 1 if $t \leq T/2$ and equal to 2 if $t > T/2$. We now give densities additional superscripts that denote the estimation sample.

2. Estimate (or calibrate) each model $m = 1, ..., M$ for each subsample s. This means that for each model we get a distribution $p_m^s(\Omega_m)$ for each subsample. This is the most time-consuming step of the algorithm, but can be easily parallelized.

3. For each model and subsample construct $p_m^{*,s}(y_{t+h,j})$ by first constructing the forecast density conditional on parameters and a given shock (see Section 2.4 for an example

[7] The pooling weights thus do not necessarily represent a ranking or evaluation of the models. Rather, the weights are chosen to optimize the performance of the *averaged* model in terms of the log-score objective function.

[8] By allowing for both Bayesian and frequentist models to enter our model pool, we implicitly equate the interpretations of uncertainty in Bayesian and frequentist frameworks. This is very much in the spirit of much applied work, which compares error bands across Bayesian and frequentist approaches, disregarding philosophical differences between the two approaches. Nevertheless, in our applications below, we do not compare across paradigms.

on how to do this in VAR models). Then average over draws from the relevant $p_m^s(\Omega_m)$ density. This step can also be parallelized.

4. Compute model weights by solving the following maximization problem for each horizon h and each variable j separately:

$$\max_{\sum_{m=1}^{M} w_{m,h}^j=1, w_{m,h}^j \geq 0} \sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_{m,h}^j p_m^{*,3-s(t)} \left(y_{t+h,j} \right) \right] \tag{3}$$

   The superscript of the density $p_m^{*,3-s(t)}$ clarifies that we use out-of-sample forecasts to construct the objective function. Geweke and Amisano (2011) provide conditions for the concavity of the objective function.

5. With model weights in hand, we can construct weighted averages of impulse responses and other statistics of interest from each model.[9]

2.4 Illustrative example: Constructing p^* for a VAR(1)

For concreteness, we now illustrate how to construct the forecasting density $p_m^* \left(y_{t+h,j} \right)$ in the context of a linear Gaussian VAR(1):

$$y_t = B y_{t-1} + u_t \tag{4}$$

$$u_t = C \varepsilon_t, \tag{5}$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$ is a vector of structural shocks and $V\left[u_t\right] = CC'$. In terms of the notation from the previous section and assuming this VAR is model 1, we have $\Omega_1 = [vec(B)' \; vec(C)']'$, where *vec* denotes columnwise vectorization of a matrix. The impulse response of y_t to shock i at horizon h is then $B^h C_{\bullet,j}$, where $C_{\bullet,j}$ is the jth column of the matrix C.

Given B and C, we can compute the on-impact conditional distributions:

$$E\left[y_t \mid y_{t-1}, \varepsilon_{t,j}\right] = B y_{t-1} + C_{\bullet,j} \varepsilon_{t,j} \tag{6}$$

$$V\left[y_t \mid y_{t-1}, \varepsilon_{t,j}\right] = CC' - C_{\bullet,j} C_{\bullet,j}' \tag{7}$$

[9] Once we have obtained the model weights, we re-estimate each model using the entire sample to obtain a final estimate of $p_m(\Omega_m)$ and use that distribution to construct our statistics of interest.

and iterate forward:

$$E\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right] = BE\left[y_{t+h-1} \mid y_{t-1}, \varepsilon_{t,j}\right] \tag{8}$$

$$V\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right] = BV\left[y_{t+h-1} \mid y_{t-1}, \varepsilon_{t,j}\right] B' + CC' \tag{9}$$

The predictive density of the vector y_t *conditional on parameters* h periods ahead is then Gaussian with conditional means and variances defined above. Furthermore, the forecasting distribution of a specific variable $y_{t,j}$ conditional on parameters and the shock is given by a normal distribution where the mean and variance are the relevant elements of $E\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right]$ and $V\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right]$.

With the internal instrument VAR, which we use in our Monte Carlos and empirical applications in Sections 3 and 4, an econometrician observes the shock if she knows the parameters. More generally, we replace $\varepsilon_{t,j}$ with $\hat{\varepsilon}_{t,j}$, the jth element of $\hat{\varepsilon}_t \equiv C^{-1}(y_t - By_{t-1})$, the fitted value of ε_t.

If B and C are estimated, we can account for parameter uncertainty by integrating over their posterior or asymptotic distribution. We implement this by averaging the predictive density across draws in a Bayesian framework, for example.

2.5 Extensions

We close our description of the methodology with a brief discussion of straightforward ways to adapt or extend our approach to several common settings.

In many scenarios, macroeconomists use identification schemes that do not point identify the structural shock of interest (such as in the case of sign restrictions in VARs). To accommodate such cases, we can enlarge the parameter vector Ω_m for any model m, where the structural shock is not point-identified, to include a parameter that selects one possible value of the structural shock consistent with the other parameters of the model. In a VAR, this would be a rotation matrix that maps the covariance matrix of the one-step ahead forecast error into the matrix of impact impulse responses. While this parameter is by definition not point identified, it does not conflict with our approach.

Similarly, it is numerically straightforward to accommodate models with nonlinearities where the models are conditionally linear and Gaussian. Key examples are VAR models with parameters that follow discrete (Sims and Zha, 2006) or continuous (Cogley and Sargent, 2005; Primiceri, 2005) Markov processes, where the respective innovations are independent of other innovations in the model. In these cases, we need to enlarge the parameter vector Ω_m to include estimates of the time t state of the Markov process.

Our approach typically gives larger weights to models that are better at forecasting the series of interest at a given horizon. This forecasting ability can be due to the inclusion of the structural shock or due to other features of each model. If a researcher wants to reward models with a larger weight when the inclusion of the structural shock *improves* forecast ability, the following alternative objective function could be used:

$$F(\phi) = \underbrace{\left[\sum_t \sum_m w_m p_m^* \left(y_{t+h,j}\right)\right]}_{\text{standard objective}} + \underbrace{\phi \left[\sum_t \sum_m w_m \left(p_m^* \left(y_{t+h,j}\right) - \overline{p_m}\left(y_{t+h,j}\right)\right)\right]}_{\text{reward for forecast improvement due to structural shock}},$$

where we define

$$\overline{p_m}\left(y_{t+h,j}\right) = \int p\left(y_{t+h,j} | y^{t-1}, \Omega_m, \mathcal{M}_m\right) p_m(\Omega_m) d\Omega_m \tag{10}$$

as the forecast density based on model $\mathcal{M}_m$ when the structural shock is not used as a predictive variable. The parameter ϕ governs how much the researcher rewards forecast improvement due to the inclusion of a structural shock. For simplicity, we set $\phi = 0$ and ignore the forecast improvement from the structural shock, as is typical in most model averaging in the literature. As a first pass to assess the viability of this modification, one could also compare weights based on $\overline{p_m}$ and p_m^*.

3 Monte Carlo Simulations

We now present three Monte Carlo exercises to illustrate our methodology. First, we consider a univariate example with two alternative models that produce consistent estimates but differ in finite sample. Second, we consider a model in which the VAR is misspecified but the LP produces consistent estimates. Third, we consider data simulated from a DSGE model, such that both the VAR and LP are misspecified. We report biases and standard deviations of our approach vis-a-vis individual models, following the common focus of the literature on first and second moments. However, our approach targets the entire forecast distribution and is not restricted to these moments.

3.1 AR(1)

As an initial proof of concept, we first consider the AR(1) Monte Carlo exercise from Herbst and Johannsen (2020). We show that in this setting, our model averaging approach performs close to optimally on a number of dimensions.

Data-Generating Process. We generate data from the univariate model:

$$y_t = \rho y_{t-1} + e_t + v_t \tag{11}$$

where $(e_t, v_t)' \overset{iid}{\sim} \mathcal{N}(0, I)$. We take $\rho = 0.97$ and use $T = 80$ observations. We seek the impulse response of y_t to a shock e_t.

Models. We estimate models of the form:

$$y_{t+h} = \beta_m^{(h)\prime} x_{m,t} + \varepsilon_{m,t+h}^{(h)}, \tag{12}$$

and use our methodology to compare the two specifications for x_t considered by Herbst and Johannsen (2020):

- **With controls:** $x_{m,t} = (e_t, y_{t-1})'$.
- **Without controls:** $x_{m,t} = e_t$.

Both specifications produce consistent estimates $\beta_{m,1}^{(h)}$ of the impulse response at horizon h. However, the second specification does not control for lagged y_t, resulting in differing finite sample performances between models. Herbst and Johannsen (2020) show that the two specifications produce different finite sample biases. The variances of the estimated impulse responses also differ.

Results. Figure 1 shows the results averaged across 5×10^4 simulations. The weights produced are intuitive and perform well on a number of dimensions. The top left panel shows that optimal weights tend to favor the model with a smaller bias. The weights are closer to 0.5 when the biases of the two models are closer.

The remaining panels show that the resulting mixture model performs well. First, the bias of the mixture model is close to the optimum that one could get with each individual model horizon by horizon. Second, the standard deviation of the mean estimate from the mixture model is also close to the lower envelope of the two individual models.[10] Third, we compute the density of the true impulse response under each of the models in the lower left panel.[11] It shows that the average density is relatively high under the averaged model with

[10] We compute the standard deviation of the impulse responses for each Monte Carlo sample and then average across all samples. Both the sample specific bias and standard deviation depend on the estimated weights for that specific sample. The averages we report in our figure for the Monte Carlo experiments thus take into account sample variation in the estimated weights.

[11] To be specific, we compute the model-specific predictive density evaluated at the true impulse response and then average across Monte Carlo repetitions.

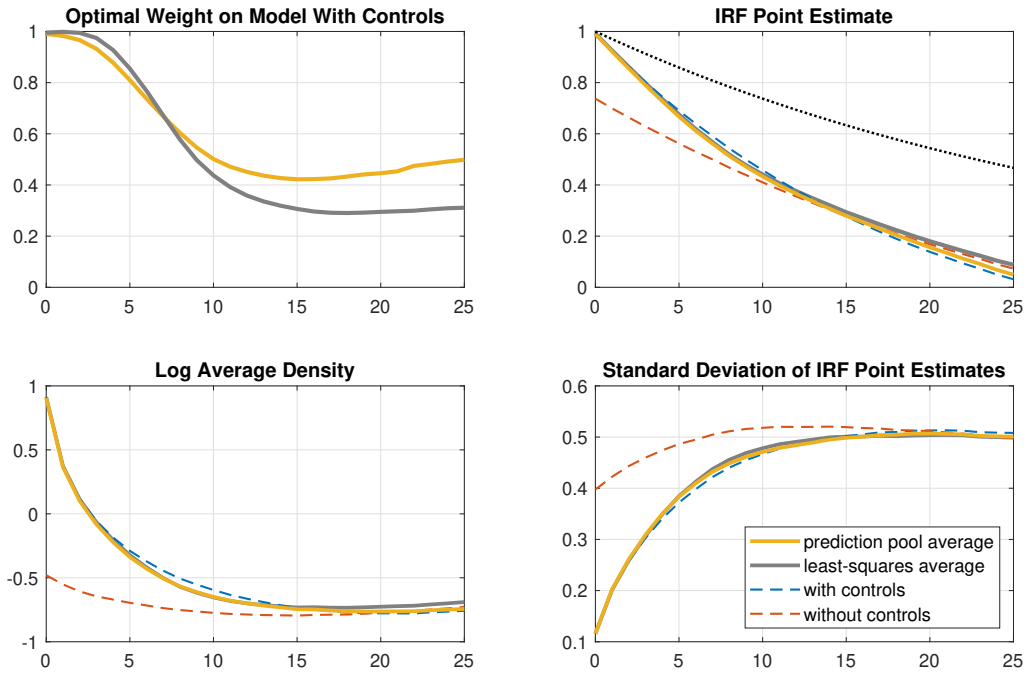

Figure 1: **Top left:** Prediction pool and least-squares weights on model with controls; **Top right:** Estimated impulse responses under each specification, averaged models, and true model; **Bottom left:** Log average probability density of true impulse response under each specification and averaged models; **Bottom right:** Standard deviation of point estimate of impulse response estimates. Dashed lines correspond to individual models, solid lines correspond to averaged models, and dotted line corresponds to true impulse response. All plots show averages across all Monte Carlo repetitions.

optimal weights.

We also compare the results to optimal weights computed using a least-squares objective function, which replaces the optimization problem in equation (3) with:

$$\min_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{t=1}^{T-h} \left(y_{t+h} - \sum_{m=1}^{M} w_m \hat{y}_{m,t+h}^{(h)} \right)^2, \tag{13}$$

where $\hat{y}_{m,t+h}^{(h)} = \beta_m^{(h)\prime} X_{m,t}$ is the fitted value of y_{t+h} in model m. We use the same sample-splitting scheme as with the prediction pool weights.

Even though the least-squares objective function targets the bias and the standard deviation of the averaged point estimates, the prediction pool performs similarly on both these measures. The prediction pool is thus able to obtain close to optimal point estimates according to this least squares objective while taking into account the entire probability distribution

for the estimated impulse response in each simulation. In situations where the forecasting density is more complicated, this is not necessarily guaranteed to be true and weights based on such a least squares objective could miss important features of the data.

3.2 Misspecified Shock

We now present an example in which the VAR is misspecified but the LP produces consistent estimates. The example highlights how the weights trade off the flexibility of the LP and the structure and relatively tighter estimates of the VAR. In addition, in finite sample the two models produce impulse responses with biases of opposite signs, offsetting each other once we average their impulse responses.

Data-Generating Process. We consider data generated from the model:

$$y_t = \rho y_{t-1} + v_{1,t} + v_{2,t} \tag{14}$$

$$v_{2,t} = \gamma v_{2,t-1} + e_{2,t}, \tag{15}$$

where

$$\begin{bmatrix} v_{1,t} \\ e_{2,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1-\gamma^2 \end{bmatrix} \right)$$

and $(\rho, \gamma) = (0.97, 0.75)$. Our parameterization ensures that the long-run variance of $v_{2,t}$ is one, which is equal to the variance of $v_{1,t}$.

We seek the impulse response of y_t to a shock $v_{1,t}$. Under the data-generating process, the impulse response at horizon h is ρ^h. We obtain results from 1000 simulations of 500 periods each.

Models. The first model we use to estimate the impulse response is an internal instrument VAR (Noh, 2018; Plagborg-Møller and Wolf, 2021):

$$\begin{bmatrix} z_t \\ y_t \end{bmatrix} = B \begin{bmatrix} z_{t-1} \\ y_{t-1} \end{bmatrix} + u_t, \tag{16}$$

where z_t is the shock of interest and u_t is assumed to be independent over time. The impulse response at horizon h is $B^h C_{\bullet,1}$, where C is the lower triangular matrix satisfying $CC' = V[u_t]$, obtained using a Cholesky decomposition.[12] As before, $C_{\bullet,1}$ is the first column

[12] This is closely related to the VARX model from Bagliano and Favero (1999) and Paul (2020), where the instrument is included as an exogenous variable in a VAR.

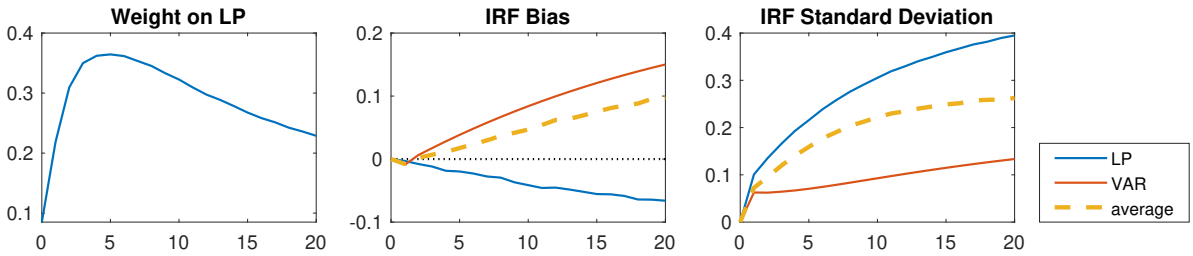

Figure 2: Prediction pool weights, biases, and asymptotic standard deviations from Monte Carlo with persistent shocks. Biases and standard deviations averaged across simulations. **Left:** Optimal weights on LP; **Middle:** Bias of impulse responses; **Right:** Standard deviation of impulse responses. All plots show averages across all Monte Carlo repetitions.

of C. We assume for simplicity that the shock is perfectly observed, i.e., $z_t = v_{1,t}$. We estimate the model equation by equation using least squares, with standard errors computed using the "wild" bootstrap (Gonçalves and Kilian, 2004).

The second model we consider is an LP:

$$y_{t+h} = \beta^{(h)} v_{1,t} + \gamma_v^{(h)} v_{1,t-1} + \gamma_y^{(h)} y_{1,t-1} + \varepsilon_{t+h}^{(h)}. \tag{17}$$

The estimated impulse response at horizon h is $\beta^{(h)}$. The model is estimated using least squares, with White standard errors (Montiel Olea and Plagborg-Møller, 2021).

The two models face a bias-variance trade-off highlighted by Li et al. (2022). The VAR (16) is misspecified because the autocorrelation of the shock u_t is assumed to be zero. This induces bias even asymptotically. The LP produces consistent estimates, with a finite sample bias that vanishes as the sample size goes to infinity. However, the structure of the VAR induces a smaller variance than the LP. Our averaging approach balances both considerations while also taking into account the finite sample performance of each method.

Results. The results are summarized in Figure 2. While the majority of the weight is placed on the VAR, there is substantial weight of up to almost 0.4 placed on the LP. The weight on the LP peaks around $h = 4$, but remains above 0.2 for all horizons after impact. By averaging the two models, we obtain an impulse response that has a lower standard deviation and only slightly larger bias than the LP. Since the VAR and LP have biases of opposite signs, averaging them tends to offset each other.[13] In this case, the difference in standard deviations leads to a larger weight on the VAR.

More generally, a correctly-specified or more flexible model need not dominate a mis-

[13]When we use weights from in-sample predictive densities instead of splitting the sample, we find that the bias almost completely vanishes. See Figure A.2 in Appendix A

specified model in finite sample. The finite sample performance of each model may not correspond to their asymptotic behavior. Furthermore, these properties may differ across impulse response horizon or the variable of interest. Our impulse response averaging approach flexibly accounts for these by constructing an optimal composite impulse response variable by variable and horizon by horizon. Figure 2 also highlights the trade-off between bias and standard deviation that is present in practically any model-averaging exercise (unless one model dominates both in terms of bias and standard deviation). Our approach reduces the bias relative to the VAR, but does so by increasing the standard deviation. However, our approach outperforms the individual models in terms of the log predictive score by construction.

3.3 Medium-Scale New Keynesian Model

We next consider a Monte Carlo exercise with data generated from a quantitative dynamic stochastic general equilibrium (DSGE) model to connect more closely to actual empirical settings. We use a DSGE model as our data-generating process because it implies VARMA (Vector Autoregressive Moving Average) dynamics for the vector of observables, so that both models we consider, VARs and LPs, are misspecified. Despite using closely related models, we find very different estimates in finite sample. The averaged impulse response balances the bias-variance trade-off, and in some cases even has a smaller bias than either individual model.

Data-Generating Process. We simulate data from the log-linearized medium-scale New Keynesian model from Smets and Wouters (2007) with parameters fixed at the posterior mode reported in the paper. We use the model to generate 150 periods of simulated data for the seven observables used by Smets and Wouters (2007) to estimate the model: GDP growth, consumption growth, investment growth, wage growth, hours, inflation, and the federal funds rate. We will focus on the impulse response of each variable to a monetary shock, which we assume to be observed by the econometrician. We obtain results across 200 simulations.

Models. We compare two models: an internal instrument VAR (16) estimated using Bayesian methods, and the Bayesian LP (Miranda-Agrippino and Ricco, 2021a). The

Bayesian LP estimates:

$$\begin{bmatrix} z_{t+h} \\ y_{t+h} \end{bmatrix} = B^{(h)} \begin{bmatrix} z_t \\ y_t \end{bmatrix} + u_{t+h}^{(h)}. \tag{18}$$

for each horizon $h > 0$. The impulse response at horizon h is $B^{(h)}C_{\bullet,1}$, where $C_{\bullet,1}$ is obtained from (16). Miranda-Agrippino and Ricco (2021a) show how to impose a prior on the model and estimate the LP impulse response analogously to a Bayesian VAR.[14] Both models have one lag and use the same Minnesota prior. In addition, we assume that the shock z_t is perfectly observed.

The two models are closely connected. First, if $B^{(h)} = B^h$, then the VAR and LP produce identical impulse responses. In particular, given the same priors, the two models would produce identical on-impact impulse responses. Next, as pointed out by Plagborg-Møller and Wolf (2021), under the appropriate regularity conditions, the two models asymptotically produce identical impulse responses. However, as the Monte Carlo exercises will show, in finite sample and under misspecification the two models can lead to substantially different estimates despite their close connections, emphasizing the need for a systematic way to average across models.[15]

Results. The weights, averaged across simulations, are summarized in the left panels of Figure 3. Overall, the prediction pools place greater weight on the VAR, with the LP typically getting a weight of 0.2 or less. The weight on the LP tends to fall at longer horizons. Nevertheless, there are non-trivial weights on the LP, especially for inflation, the federal funds rate, and hours.

The middle and right panels of Figure 3 plot the average biases and standard deviations of the impulse response functions, providing an explanation for the small weights on the LP. First, even though the LP has greater flexibility, in many cases its bias tends to be larger or is at most of similar magnitude relative to the VAR. This arises partly due to the relatively short sample of 150 periods. Second, the right panels show that the LP has substantially larger posterior standard deviation, which is consistent with evidence reported in the literature (Miranda-Agrippino and Ricco, 2021b; Li et al., 2022). The difference

[14] Miranda-Agrippino and Ricco (2021a) do not explicitly model autocorrelation in the residual of their LP specification, but instead use a sandwich-type estimator for the posterior covariance matrix.

[15] There are two differences of note relative to Plagborg-Møller and Wolf (2021). First, because we impose a prior, the estimated impulse responses at horizon $h > 0$ differ even if the least squares estimates are equivalent. In particular, for longer horizons, the likelihood of the LP becomes more dispersed, bringing the posterior closer to the prior. Second, the estimated system (18) differs from the LP setup used in Plagborg-Møller and Wolf (2021).

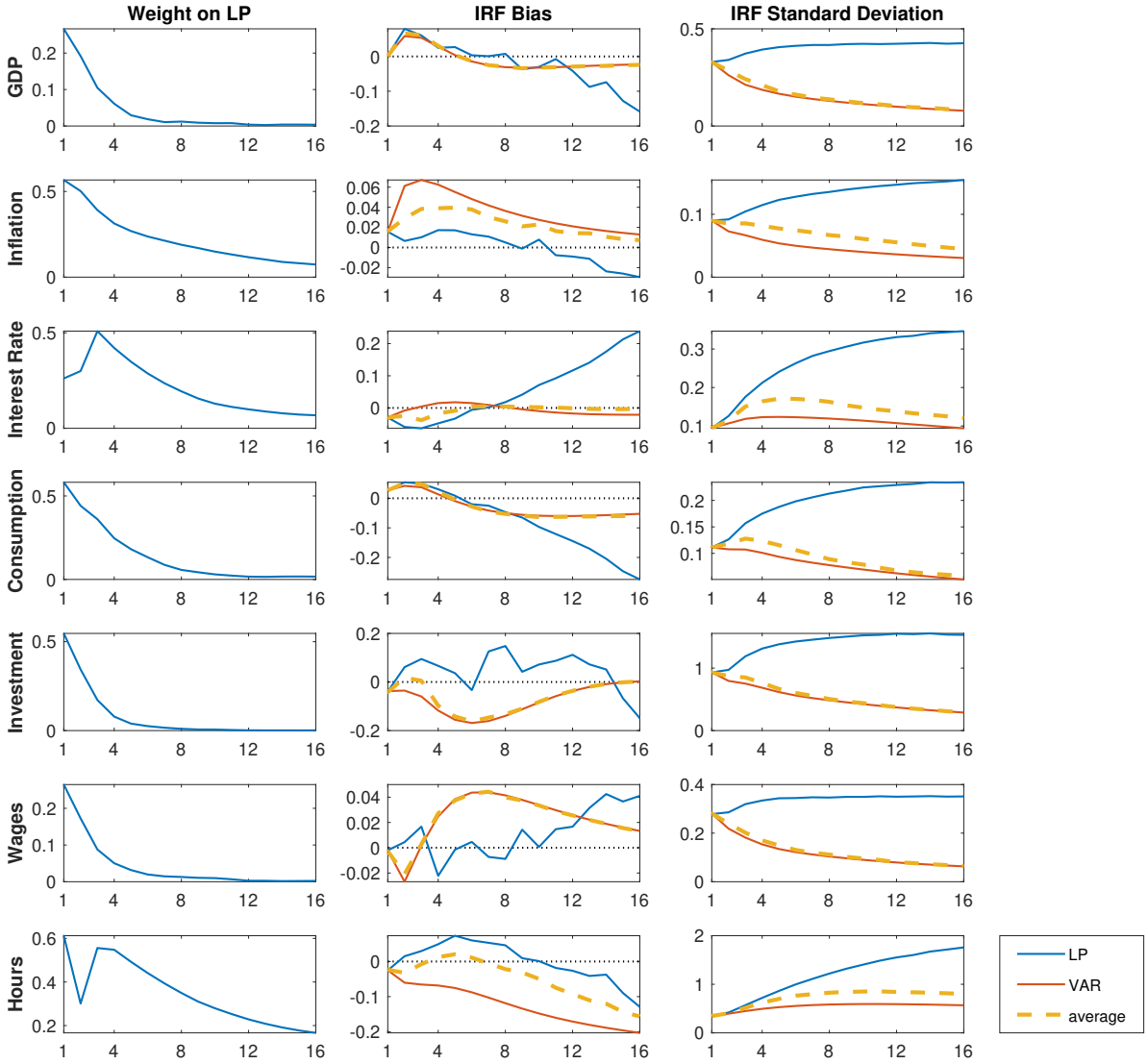

Figure 3: Prediction pool weights, biases, and posterior standard deviations from Smets and Wouters (2007) Monte Carlo. Biases and standard deviations averaged across simulations. **Left:** Optimal weights on LP; **Middle:** Bias of impulse responses; **Right:** Posterior standard deviation of impulse responses. All plots show averages across all Monte Carlo repetitions.

in standard deviations is especially large at longer horizons, which accounts for the lower weights on the LP at those horizons.

To get a better sense of what is driving the weights, we focus first on the impulse response for inflation. The weights on the LP are relatively high, especially in the initial quarters. Correspondingly, we find that the LP estimates a response that has a smaller bias than the VAR. Nevertheless, the VAR continues to receive more than half the weight because its impulse response standard deviation is smaller. At longer horizons, the weight on the LP

falls as the difference in biases shrinks while the difference in standard deviations increases.

The impulse response for hours further illustrates the behavior of the prediction pools. The weights on the LP increases over the first four quarters but does not decay as quickly as for other variables. Even though the standard deviations are similar initially, the LP displays a substantially larger bias than the VAR at short horizons, reducing its optimal weight. Subsequently, the LP and VAR have biases of opposite signs that offset each other when averaged, as was the case in our previous Monte Carlo exercise. By averaging the impulse responses, the prediction pool can produce an average impulse response that has a smaller bias than either model, with the bias almost completely eliminated at horizon $h = 15$. At longer horizons, the weights trade off two forces. First, the VAR bias begins to increase while the LP bias begins to decrease. Second, the LP posterior standard deviation increases while the VAR standard deviation remains relatively constant. In balance, the weights begin to favor the LP less at longer horizons, but with a decline that is less steep than in other variables.

Overall, the results here emphasize two key messages. First, the relative biases and variances of the models differ depending on variable and horizon. Prediction pools offer the flexibility to trade off these properties variable by variable and horizon by horizon, thus making full use of the relative strengths of each model. Second, even when models have similar asymptotic properties, there can be substantial gains from averaging over them in finite sample. In particular, the bias of the average impulse response can in some cases be lower than that of either individual model.

4 Empirical Applications

We now apply our methodology to estimating impulse responses to monetary shocks on actual data. We consider two identification schemes, where we use external instruments that have recently become popular in the macroeconomic literature, namely a high-frequency identification instrument and a narrative instrument. These instruments have featured prominently in many LP applications, but have recently also been used in VARs. In both applications we consider a range of plausible empirical models in our prediction pools. Overall, our applications indicate that prediction pools offer a more plausibly accurate assessment of the dynamic effects of monetary shocks as they optimally resolve the bias-variance trade-off, especially when, as is likely, the underlying models are misspecified.

4.1 High-Frequency Identification

The first empirical application uses a monthly VAR to study the effects of a monetary shock using the high-frequency identification instrument from Gertler and Karadi (2015), similar to Caldara and Herbst (2019). In particular, we use data on industrial production (IP), unemployment, the producer price index for finished goods, the federal funds rate, and the Baa corporate bond spread. We consider a sample with monthly data from March 1990 through November 2007. As in the Smets and Wouters (2007) Monte Carlo exercise, we consider two models: a Bayesian internal instruments VAR and a Bayesian LP. We choose 12 lags for both models and use a Minnesota prior. The tightness parameter for the VAR is selected to maximize the marginal likelihood, while the tightness parameter for the LP is fixed at 1.[16]

Results. The empirical application results, shown in Figure 4, illustrate the importance of computing the weights variable by variable and horizon by horizon, rather than having a single weight for all variables and horizons. In particular, while almost all the weight for the IP impulse response is placed on the VAR, the prediction pool places majority of the weight on the LP at some horizons for each of the other variables. In addition, most of the variables feature weights on the LP that range from zero to one depending on the horizon. These results extends the typical finding that relative forecast performances depend on variable and horizon to our particular interpretation of impulse responses and conditional forecasts.

The large weights attached to the LP for certain variables and horizons contrasts with the Smets and Wouters (2007) Monte Carlo, where the average weights were never much above 0.5. This reiterates the message that the relative performance of each model depends on the setting, including the data-generating process and sample size.

The impulse response of unemployment provides one example of how the averaged impulse response may be qualitatively different than each individual model. The prediction pool places approximately equal weight on the LP and VAR after the three-year horizon. The resulting averaged impulse response has a peak response from unemployment approximately two-and-a-half years after the shock that subsequently reverts to zero, in contrast to the persistent negative and positive responses from the LP and VAR, respectively.

There are also cases where we find intuitive reasons for more uneven weights. For example, in the IP impulse response, the prediction pool places almost all the weight on the VAR. The average impulse response thus displays a persistent decline in the level of IP in response

[16]The prior for the LP is chosen to be flatter than in the VAR because the likelihood for the LP is dominated by a Minnesota prior with tightness parameter matching the VAR. A tightness parameter of 1 provides some shrinkage while allowing the data to speak.

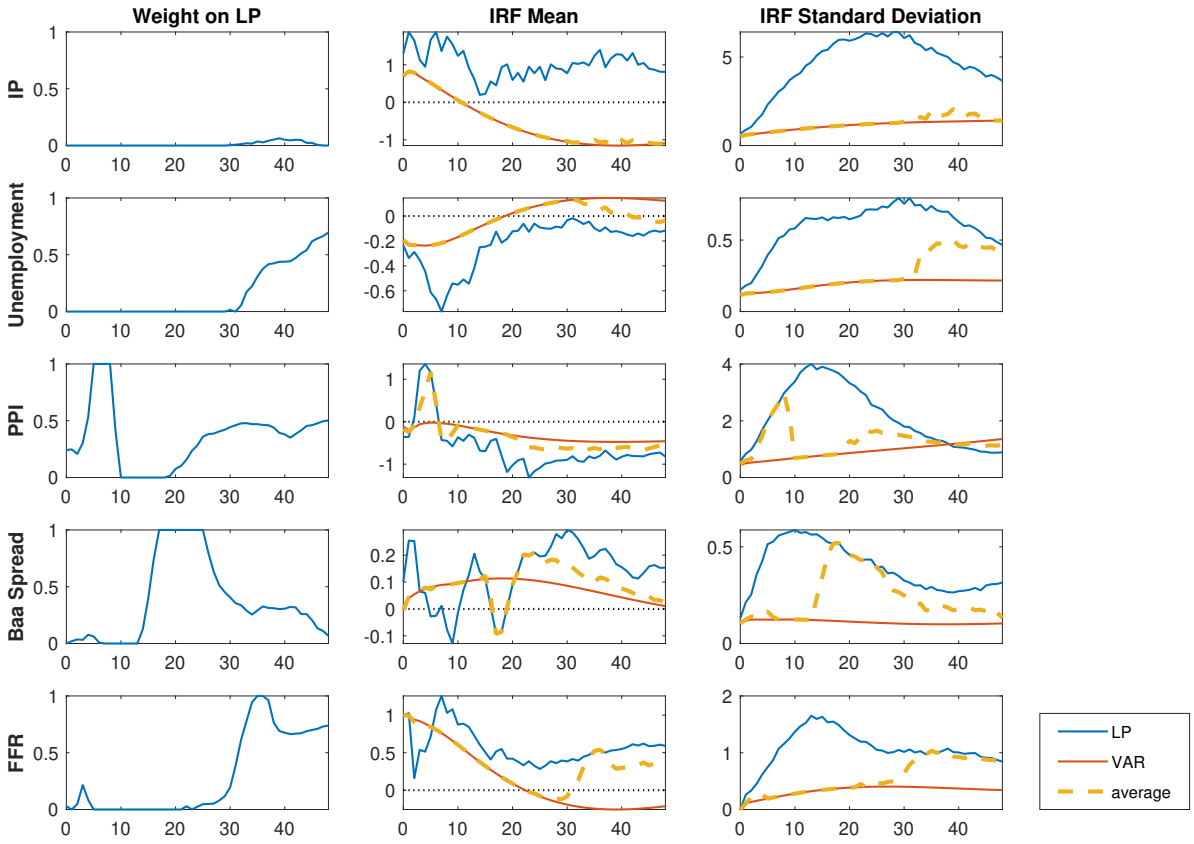

Figure 4: Prediction pool weights, posterior mean, and posterior standard deviation from high-frequency identification empirical application. **Left:** Optimal weights on LP; **Middle:** Posterior mean of impulse responses; **Right:** Posterior standard deviation of impulse responses.

to a contractionary monetary shock. This is closer to predictions from economic theory than the persistent positive response that the LP finds.

An intermediate case is the impulse response for the federal funds rate at long horizons. Here the LP is favored but the VAR continues to receive a nontrivial weight. The resulting point estimate is positive, but with a large standard deviation. Once again, the averaged impulse response is arguably more reasonable than each individual response. Relative to the LP, the averaged impulse response has a mean that is slightly closer to zero. Unlike the VAR that produced a relatively tightly estimated negative response, the averaged response has a large standard deviation.

As in the Monte Carlo exercises, we see benefits to averaging the impulse responses using prediction pools. The averaged impulse responses are more plausible than either the LP or the VAR. Even more than the Monte Carlos, the flexibility of the prediction pools is critical, with the weights varying dramatically across variables and horizons.

4.2 Narrative Instrument

Our second empirical application follows the study of the Romer and Romer (2004) shocks in Ramey (2016). In particular, we use monthly data on the log of IP, the unemployment rate, the log of the CPI, the log of a commodity price index, the federal funds rate, and the Romer and Romer (2004) instrument for March 1969 through December 1996.

We consider four models, each estimated using frequentist methods:

1. **Cholesky VAR.** Following Coibion (2012), we estimate a VAR with the log of IP, the unemployment rate, the log of the CPI, and the log of the commodity price index in the first block, followed by the cumulated Romer and Romer (2004) instrument ordered last. The monetary shock is assumed to be the last shock from a Cholesky decomposition.

2. **Internal Instrument VAR.** We estimate a VAR with the Romer and Romer (2004) instrument as the first variable, followed by the log of IP, the unemployment rate, the log of the CPI, the log of the commodity price index, and the federal funds rate. The monetary shock is assumed to be the first shock from a Cholesky decomposition.

3. **LP With Recursiveness Assumption.** We follow Ramey (2016) and estimate regressions of the form

$$z_{t+h} = \alpha_h + \theta_h \cdot \text{shock}_t + \text{control variables} + \varepsilon_{t+h}, \tag{19}$$

   where z_{t+h} is the variable of interest, shock_t is the Romer and Romer (2004) instrument. The control variables include lags of the Romer and Romer (2004) shock, the log of IP, the unemployment rate, the log of the CPI, the log of the commodity price index, and the federal funds rate, as well as contemporaneous values of the log of IP, the unemployment rate, and the log price indices to preserve the recursiveness assumption, as in the Cholesky VAR.

4. **LP Without Recursiveness Assumption.** This is identical to the LP with the recursiveness assumption, except that we do not control for contemporaneous variables. This makes the assumption that the Greenbook forecasts used by Romer and Romer (2004) already include all information used by the Fed for setting interest rates.

Following Ramey (2016), both VARs use twelve lags and both LPs use two lags. In principle, all four models estimate the same impulse response using the same instrument. However, the models make different identifying assumptions, include different controls, and have different

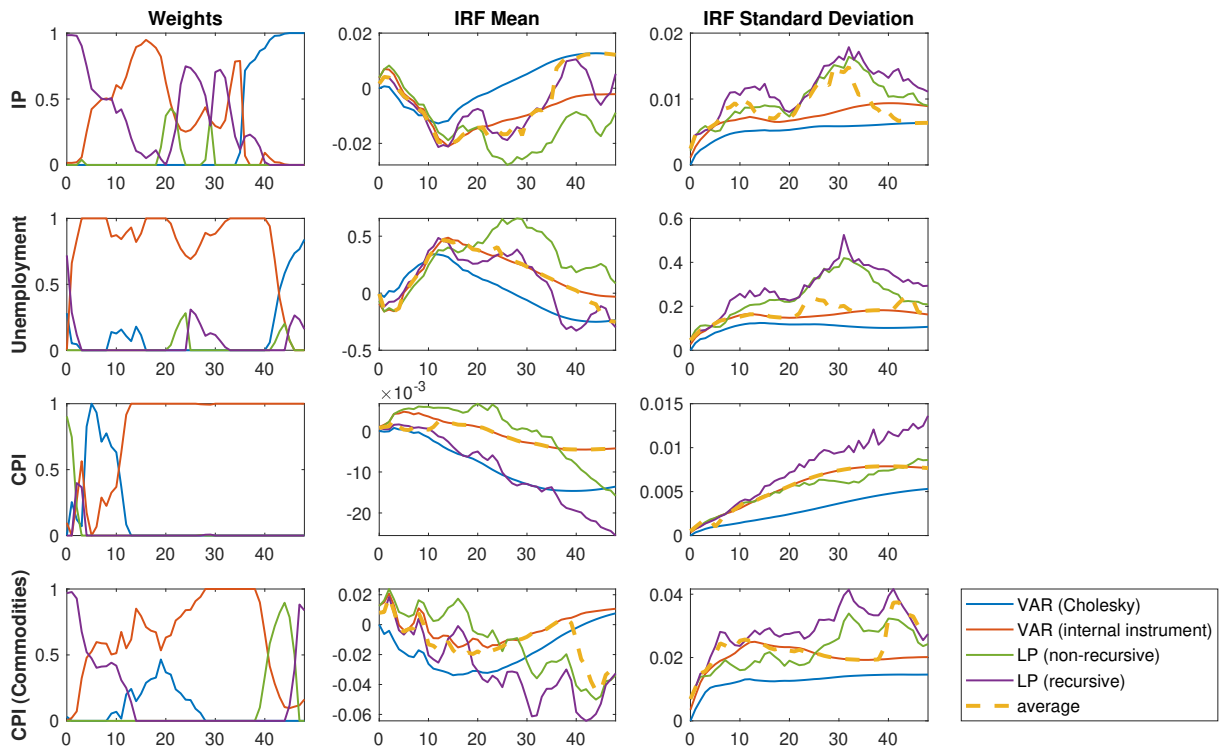

Figure 5: Prediction pool weights, posterior mean, and posterior standard deviation from narrative identification empirical application. **Left:** Optimal weights on LP; **Middle:** Posterior mean of impulse responses; **Right:** Posterior standard deviation of impulse responses.

numbers of lags. Importantly, this means that unlike the previous application, the LPs do not nest the VARs here.

Results. The results are summarized in Figure 5. In general, majority of weight is placed on the LPs on impact, while the VARs get assigned greater weight at longer horizons. This is similar to the Smets and Wouters (2007) Monte Carlo and differs from the high-frequency instrument application in Section 4.1. However, unlike the Smets and Wouters (2007) Monte Carlo, the relative weights are arguably not tightly connected with the respective standard deviations. In particular, the internal instrument VAR receives a large weight even though it generally has a higher variance than the Cholesky VAR and a similar variance to the LP without the recursiveness assumption.

The impulse response for IP yields two interesting observations. First, at the one- to three-year horizon, the prediction pool places majority of the weight first on the internal instrument VAR then the recursive LP, yielding a deeper and more prolonged contraction in response to the identified shock than implied by the Cholesky VAR. The fact that the latter is not favored despite its lower variance suggests that the average impulse response

provides a better fit to the data. However, after three years, the Cholesky VAR becomes heavily weighted, implying a rebound in IP rather than a convergence back to trend.

The weights for the unemployment and CPI impulse responses are mostly on the internal instrument VAR. One notable exception is unemployment at the long horizon. Like IP, the weights favor the Cholesky VAR and are associated with a rebound in economic activity, with unemployment undershooting its trend.

The average impulse response for commodity prices is fairly noisy, with substantial weight placed on the LPs after the three-year horizon. This results in a high variance at those horizons. In particular, the variance of the averaged impulse response is higher than any of the individual responses at the four-year horizon. Given the volatility of commodity prices, it is plausible that the data are not informative about their response to the identified shock. As a result an impulse response with high variance may fit the data best from the standpoint of the log predictive score function, which accounts for the full predictive density rather than just the point estimates.

The above results stress the fact that prediction pools utilize the predictive density and not only particular moments of the estimated impulse responses. While tighter estimates may reduce mean-square error, they need not increase the log predictive score function (3). The prediction pool not only favors models whose fit to the data sufficiently dominate the rest of the pool, but may even favor a large variance if prediction is, in fact, difficult.

As a final exercise, we drop the Cholesky VAR from the set of models and repeat the exercise. The results are shown in Figure 6. The main observation is that the weight that was previously on the Cholesky VAR primarily gets transferred to the internal instrument VAR, which is the closest model. The averaged impulse responses remain relatively similar to the exercise with all four models, illustrating the consistency of the weights assigned as we change the set of models (see Appendix B for details).

5 Conclusion

In this paper, we develop a methodology that delivers an encompassing approach to computing dynamic responses of macroeconomic variables to shocks. A fundamental challenge for empirical researchers is the choice of a specific model for such an exercise since it is well known that they come with their own issues such as bias or large standard errors. Based on the idea of prediction pools, as in Geweke and Amisano (2011), we show how to average across impulse responses from multiple models. Our framework thus presents an approach to incorporate evidence from a variety of models in a consistent and plausible manner to get closer to the actual truth in real economic data.

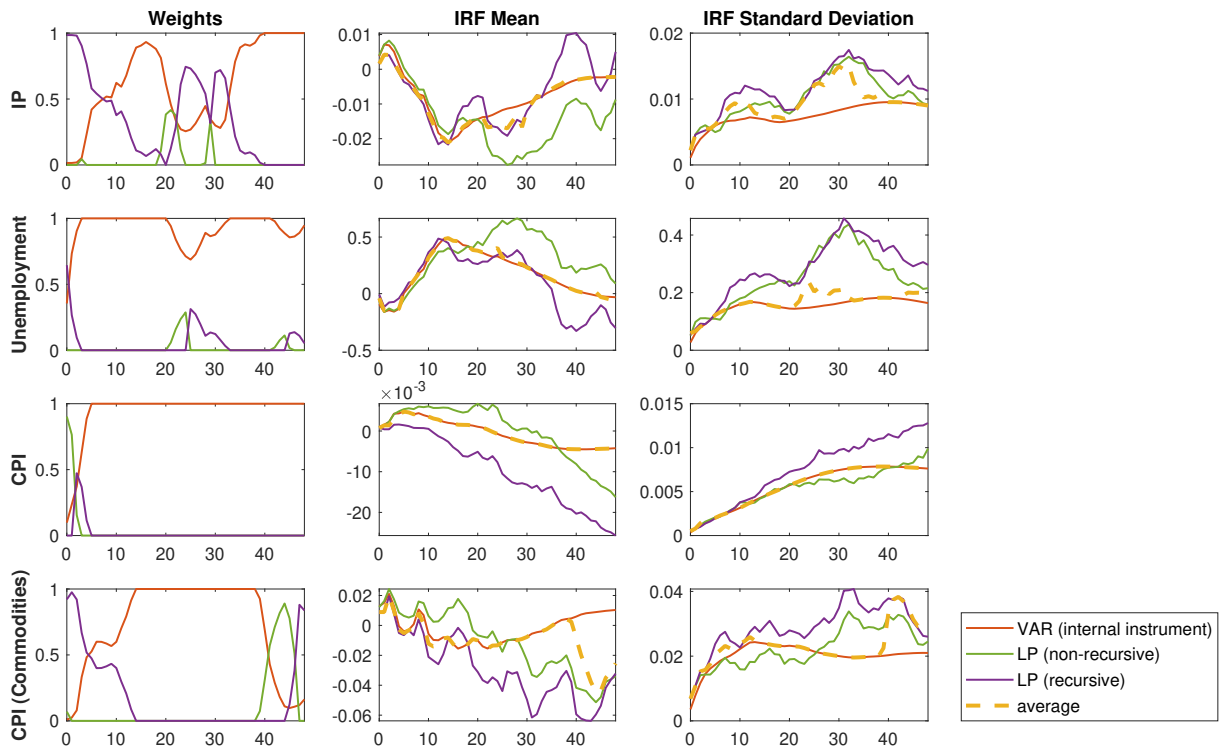

Figure 6: Prediction pool weights, posterior mean, and posterior standard deviation from narrative identification empirical application. **Left:** Optimal weights on LP; **Middle:** Posterior mean of impulse responses; **Right:** Posterior standard deviation of impulse responses.

General theorems about which class of models should be used in empirical macroeconomics are hard to come by once we make realistic assumptions. This has led to many alternative impulse response estimators coexisting in the literature. We exploit that each of these can be useful in particular situations, making empirical macroeconomics an ideal setting to exploit flexible model-averaging schemes. Our prediction-pool based approach makes model-averaging in these scenarios possible.

The key differences relative to existing methods are (i) much greater flexibility, which allows us to use, for example, estimators based on different statistical paradigms; (ii) relative small computational burden especially since the most time-consuming task can be parallelized; and (iii) ability to exploit each method's strength as much as possible by computing horizon- and variable-specific weights based on predictive densities instead of a few selected statistics.

Overall, our Monte Carlos and empirical applications highlight several broad messages for estimating impulse responses:

1. The optimal weights on models depend on horizon, variable, and application.

2. The choice of model depends on the entire predictive distribution, not only the point estimates. Our examples focus on the mean and variance, but skewness, kurtosis, or any other property of the predictive distribution could be important more generally.

3. Theoretical properties of individual models are not sufficient criteria for the choice of weights. For instance, misspecified models may dominate correctly specified (or more flexible) models in finite sample. On the other hand, models that produce tighter estimates need not receive greater weight.

Our use of prediction pools provides a systematic and computationally tractable way to account for these issues in a wide range of applications.

The flexibility of our methodology raises the possibility of numerous applications beyond what we have explored here. For example, our approach can also be extended to average across different identification schemes for the same shock (for example, studying the effects of monetary policy shocks using a VAR identified with sign restrictions and a VAR identified using an instrument). Furthermore, one could envision using our approach to discriminate between various equilibrium models that encode different transmission mechanisms for the shock of interest.

References

Amisano, Gianni and John Geweke (2017), "Prediction Using Several Macroeconomic Models." *Review of Economics and Statistics*, 99, 912–925.

Baek, ChaeWon and Byoungchan Lee (2022), "A Guide to Autoregressive Distributed Lag Models for Impulse Response Estimations." *Oxford Bulletin of Economics and Statistics*.

Bagliano, Fabio C and Carlo A Favero (1999), "Information from Financial Markets and VAR Measures of Monetary Policy." *European Economic Review*, 43, 825–837.

Barnichon, Regis and Christian Brownlees (2019), "Impulse Response Estimation by Smooth Local Projections." *Review of Economics and Statistics*, 101, 522–530.

Bates, J. M. and C. W. J. Granger (1969), "The Combination of Forecasts." *Journal of the Operational Research Society*, 20, 451–468.

Bruns, Martin and Helmut Lütkepohl (2022), "Comparison of Local Projection Estimators for Proxy Vector Autoregressions." *Journal of Economic Dynamics and Control*, 134, 104277.

Caldara, Dario and Edward Herbst (2019), "Monetary Policy, Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs." *American Economic Journal: Macroeconomics*, 11, 157–92.

Canova, Fabio and Christian Matthes (2021), "A Composite Likelihood Approach for Dynamic Structural Models." *Economic Journal*, 131, 2447–2477.

Cogley, Timothy and Thomas J. Sargent (2005), "Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S." *Review of Economic Dynamics*, 8, 262–302.

Coibion, Olivier (2012), "Are the Effects of Monetary Policy Shocks Big or Small?" *American Economic Journal: Macroeconomics*, 4, 1–32.

Del Negro, Marco, Raiden B. Hasegawa, and Frank Schorfheide (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance." *Journal of Econometrics*, 192, 391–405.

Del Negro, Marco and Frank Schorfheide (2004), "Priors from General Equilibrium Models for VARs." *International Economic Review*, 45, 643–673.

Doan, Thomas, Robert Litterman, and Christopher Sims (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions." *Econometric Reviews*, 3, 1–100.

Gertler, Mark and Peter Karadi (2015), "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics*, 7, 44–76.

Geweke, John and Gianni Amisano (2011), "Optimal Prediction Pools." *Journal of Econometrics*, 164, 130–141.

Geweke, John and Gianni Amisano (2012), "Prediction with Misspecified Models." *American Economic Review: Papers & Proceedings*, 102, 482–86.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2015), "Prior Selection for Vector Autoregressions." *Review of Economics and Statistics*, 97, 436–451.

Gonçalves, Sılvia and Lutz Kilian (2004), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form." *Journal of Econometrics*, 123, 89–120.

Hall, Stephen G. and James Mitchell (2007), "Combining Density Forecasts." *International Journal of Forecasting*, 23, 1–13.

Hansen, Bruce E. (2007), "Least Squares Model Averaging." *Econometrica*, 75, 1175–1189.

Hansen, Bruce E (2016), "Stein Combination Shrinkage for Vector Autoregressions." Working paper, University of Wisconsin, Madison.

Herbst, Edward and Benjamin K. Johannsen (2020), "Bias in Local Projections." Finance and Economics Discussion Series 2020-010, Washington: Board of Governors of the Federal Reserve System.

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

Kilian, Lutz (1998), "Small-Sample Confidence Intervals for Impulse Response Functions." *Review of Economics and Statistics*, 80, 218–230.

Li, Dake, Mikkel Plagborg-Møller, and Christian K. Wolf (2022), "Local Projections vs. VARs: Lessons From Thousands of DGPs." Working paper.

Lusompa, Amaze (2021), "Local Projections, Autocorrelation, and Efficiency." Federal Reserve Bank of Kansas City Working Paper 21-01.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series." *Journal of Econometrics*, 135, 499–526.

Miranda-Agrippino, Silvia and Giovanni Ricco (2021a), "Bayesian Local Projections." Working paper.

Miranda-Agrippino, Silvia and Giovanni Ricco (2021b), "The Transmission of Monetary Policy Shocks." *American Economic Journal: Macroeconomics*, 13, 74–107.

Montiel Olea, José Luis and Mikkel Plagborg-Møller (2021), "Local Projection Inference is Simpler and More Robust Than You Think." *Econometrica*, 89, 1789–1823.

Noh, Eul (2018), "Impulse-Response Analysis with Proxy Variables." Working paper, University of California San Diego.

Paul, Pascal (2020), "The Time-Varying Effect of Monetary Policy on Asset Prices." *Review of Economics and Statistics*, 102, 690–704.

Pesavento, Elena and Barbara Rossi (2006), "Small-Sample Confidence Intervals for Multivariate Impulse Response Functions at Long Horizons." *Journal of Applied Econometrics*, 21, 1135–1155.

Plagborg-Møller, Mikkel and Christian K. Wolf (2021), "Local Projections and VARs Estimate the Same Impulse Responses." *Econometrica*, 89, 955–980.

Primiceri, Giorgio E. (2005), "Time Varying Structural Vector Autoregressions and Monetary Policy." *Review of Economic Studies*, 72, 821–852.

Qu, Zhongjun (2018), "A Composite Likelihood Framework for Analyzing Singular DSGE Models." *The Review of Economics and Statistics*, 100, 916–932.

Ramey, Valerie A. (2016), "Macroeconomic Shocks and Their Propagation." *Handbook of Macroeconomics*, 2, 71–162.

Romer, Christina D and David H Romer (2004), "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94, 1055–1084.

Sims, Christopher A (1980), "Macroeconomics and Reality." *Econometrica*, 1–48.

Sims, Christopher A and Tao Zha (1999), "Error Bands for Impulse Responses." *Econometrica*, 67, 1113–1155.

Sims, Christopher A. and Tao Zha (2006), "Were There Regime Switches in U.S. Monetary Policy?" *American Economic Review*, 96, 54–81.

Smets, Frank and Rafael Wouters (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE approach." *American Economic Review*, 97, 586–606.

Stock, James H. and Mark W. Watson (2018), "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *The Economic Journal*, 128, 917–948.

Stock, J.H. and M.W. Watson (2016), "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In *Handbook of Macroeconomics* (J. B. Taylor and Harald Uhlig, eds.), volume 2, 415–525, Elsevier.

Stone, M. (1961), "The Opinion Pool." *The Annals of Mathematical Statistics*, 32, 1339 – 1342.

Strachan, R.W. and H.K. van Dijk (2007), "Bayesian Model Averaging in Vector Autoregressive Processes With an Investigation of Stability of the US Great Ratios and Risk of a Liquidity Trap in the USA, UK and Japan." Econometric Institute Research Papers EI 2007-11, Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.

Waggoner, Daniel F. and Tao Zha (2012), "Confronting Model Misspecification in Macroeconomics." *Journal of Econometrics*, 171, 167–184.

A Supplementary Figures

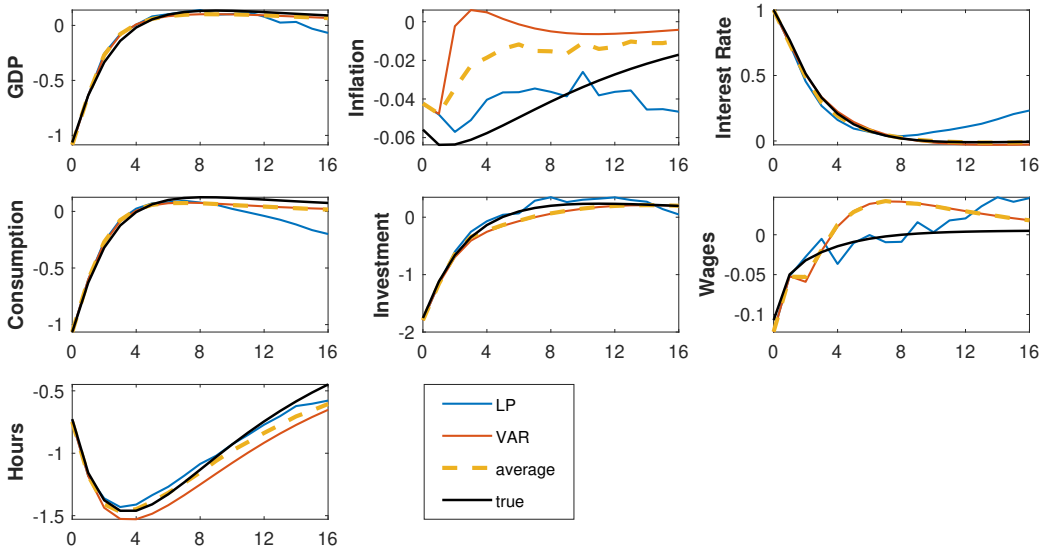

Figure A.1: Posterior mean estimates of impulse response to monetary policy shock in New Keynesian model Monte Carlo, average across simulations. GDP, consumption, investment, and wages in growth rates.

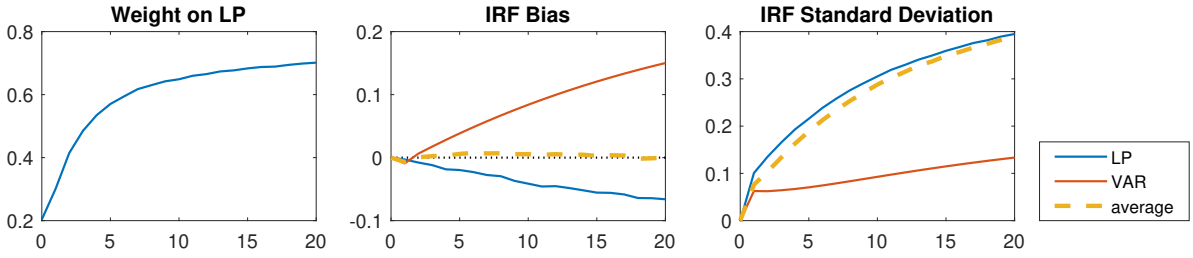

Figure A.2: Prediction pool weights, biases, and asymptotic standard deviations from Monte Carlo with persistent shocks, without sample splitting. Biases and standard deviations averaged across simulations. **Left:** Optimal weights on LP; **Middle:** Bias of impulse responses; **Right:** Standard deviation of impulse responses.

B Theoretical Results of Geweke and Amisano (2011)

In this section we state major theoretical results from Geweke and Amisano (2011) for the sake of completeness. Let us first explicitly state our objective function:

$$f_T(\mathbf{w}_h) = \max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_{m,h} p_m^* \left(y_{t+h,j} \right) \right] \tag{A.1}$$

where $\mathbf{w}_h = [w_{1,h}\ w_{2,h} \cdots w_{M,h}]'$ is the vector of model weights for a given horizon h (and a given variable j, which we have not made explicit in this notation). We will generally assume that $f_T(\mathbf{w}_h)$ is concave, i.e., $\partial^2 f_T / \partial \mathbf{w}_h \partial \mathbf{w}_h'$ is negative definite. For the case of two models ($M = 2$), Geweke and Amisano (2011) show that that the objective function will be concave if the expected difference between the two predictive densities will not be zero as the same size increases.[17] We will call a subset of models *dominant* if its weights sum to 1. A subset of models is *excluded* if each of those models has a weight of 0. With the assumption of concavity, Geweke and Amisano (2011) show the following results:

1. If $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ then $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates $\{\mathcal{M}_1, \ldots, \mathcal{M}_m, \mathcal{M}_{j_1}, \ldots, \mathcal{M}_{j_k}\}$ for all $\{j_1, \ldots, j_k\} \subseteq \{m+1, \ldots, n\}$.

2. If $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates all pools $\{\mathcal{M}_1, \ldots, \mathcal{M}_m, \mathcal{M}_j\}$ $(j = m+1, \ldots, n)$ then $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$.

3. The set of models $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ is excluded in the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ if and only if $\mathcal{M}_j$ is excluded in each of the pools $\{\mathcal{M}_j, \mathcal{M}_{m+1}, \ldots, \mathcal{M}_n\}$ $(j = 1, \ldots, m)$.

4. If the model $\mathcal{M}_1$ is excluded in all pools $(\mathcal{M}_1, \mathcal{M}_i)$ $(i = 2, \ldots, n)$ then $\mathcal{M}_1$ is excluded in the pool $(\mathcal{M}_1, \ldots, \mathcal{M}_n)$.

[17] Note that even in the case of LPs and VARs with the same right-hand side variables, it is unlikely (at least at larger horizons) that the implied predictive densities are the same even though the VAR specification is nested in LPs.