Geneva Graduate Institute (IHEID)

Topics in Econometrics

Term Paper

The Incremental Predictive Power of Consumer Sentiment in Macroeconomic Forecasting

Evidence from a Hierarchical Bayesian VAR and Forecast-Revision Diagnostics

Jingle Fu

Professor: Marko Mlikota

Abstract

Does consumer sentiment add predictive content for inflation and real activity once macro aggregates and financial prices are already included? I answer with nested information sets in a hierarchical Bayesian VAR, recursive pseudo out-of-sample forecasting, and a revision-based diagnostic of forecast updating following Coibion and Gorodnichenko (2015). Point-forecast evidence indicates limited incremental accuracy from sentiment beyond prices (Table 1; Figure 1). The revision diagnostic indicates that updating patterns vary with the information set, providing evidence about internal forecast discipline rather than behavioral beliefs (Table 2). Throughout, coefficients and shrinkage are interpreted as properties of a regularized forecasting system, not causal statements about expectations.

Keywords: Bayesian VAR; hierarchical shrinkage; forecasting; consumer sentiment; forecast revisions.

1 Introduction

This paper asks whether consumer sentiment adds incremental predictive content for inflation and real activity once conventional macro aggregates and financial prices are already included. The analysis separates *forecast accuracy* from *forecast discipline* to distinguish incremental information content from internal updating patterns in a regularized forecasting system.

Contributions and headline evidence. The paper compares nested information sets in a hierarchical BVAR with data-driven overall shrinkage, allowing dimensional changes without ad hoc tuning (Bańbura, Giannone, & Reichlin, 2010; Giannone, Lenza, & Primiceri, 2015; Kuschnig & Vashold, 2021). It separates point-forecast accuracy from revision-based diagnostics: the accuracy evidence offers limited incremental support for sentiment once financial prices are included (Table 1; Figure 1), while revision diagnostics document updating patterns that vary with the information set (Table 2). Because the specifications are nested, equal-accuracy tests are treated as suggestive and nested-robust adjustments are reported in the appendix (Clark & McCracken, 2001; Clark & West, 2007). Applied to model-implied forecasts, the Coibion and Gorodnichenko (2015) regression is used as a diagnostic rather than a behavioral claim.

Related literature. The revision diagnostic builds on Coibion and Gorodnichenko (2015) and connects to work on expectations updating, including models of diagnostic expectations; here, the regression is used as a model diagnostic rather than a structural test (Bordalo, Gennaioli, & Shleifer, 2018, 2020). Evidence on whether confidence or sentiment contains incremental forecasting information is mixed once other indicators are included, and real-time evaluations often find limited or unstable gains (Bram & Ludvigson, 1998; Carroll, Fuhrer, & Wilcox, 1994; Croushore, 2005; Ludvigson, 2004). The inflation-forecasting literature emphasizes that parsimonious benchmarks can be difficult to beat and that forecasting relationships shift, motivating cautious interpretation of small differences (Atkeson & Ohanian, 2001; Stock & Watson, 2007). Hierarchical BVAR shrinkage provides a disciplined way to compare information sets of different dimensions without ad hoc tuning, which is central to the design here (Bańbura et al., 2010; Giannone et al., 2015; Kuschnig & Vashold, 2021).

I interpret the predictive power of sentiment through the lens of a signal extraction problem. Financial prices aggregate dispersed information efficiently, potentially acting as a sufficient statistic for future macroeconomic fundamentals. In contrast, consumer sentiment surveys are characterized by idiosyncratic noise and potential measurement errors. This paper tests whether, conditional on efficient price discovery, the marginal signal content in noisy survey measures is statistically significant for point forecasts.

Roadmap. Section 2 describes the nested information sets and the evaluation setup, Section 3 presents the forecasting system and diagnostics, Section 4 reports the evidence, and Section 5 concludes.

Empirical focus. The empirical goal is descriptive: quantify the incremental predictive content of sentiment conditional on macro aggregates and financial prices, and summarize forecast-updating patterns using the revision diagnostic. I avoid structural interpretations of revision-regression coefficients and treat formal comparisons under nesting cautiously. This framing separates incremental information content from any belief-distortion interpretation.

2 Data

This section defines the nested information sets and the evaluation scale used in the forecasting comparison. The dataset combines macro aggregates, financial prices, and a survey-based measure of consumer sentiment. The information sets are nested to isolate incremental information content. The 'Small' set includes core macro variables: Industrial Production (INDPRO), Consumer Price Index (CPIAUCSL), Unemployment Rate (UNRATE), and the Federal Funds Rate (FEDFUNDS). The 'Medium' set adds financial prices: the 10-Year Treasury Yield (GS10), the S&P 500 Index (SP500), and WTI crude oil prices (DCOILWTICO). The 'Full' set adds the University of Michigan Consumer Sentiment Index (UMCSENT). The sample period covers 1985M1 - 2019M12. The comparison between Medium and Full therefore targets whether sentiment contributes beyond information already summarized in market prices.

Following standard BVAR practice, the model is estimated in levels or log-levels (Giannone et al., 2015; Sims, 1980). Forecasts are evaluated on a common growth-rate scale constructed from model-implied level forecasts, using the same transformation throughout the forecasting system. This structure isolates the incremental role of sentiment while keeping the evaluation scale consistent across information sets. For log-level series (CPIAUCSL and INDPRO), the evaluation scale is the cumulative annualized growth rate from the forecast origin,

$$g_{t,h} = \frac{1200}{h} \left(\ell_{t+h} - \ell_t \right), \qquad h \in \{1, 3, 12\},$$

so that $h = 1$ uses $1200(\ell_{t+1} - \ell_t)$, $h = 3$ uses $400(\ell_{t+3} - \ell_t)$, and $h = 12$ uses $100(\ell_{t+12} - \ell_t)$; levels are handled as cumulative changes per period. This definition matches the code-based transformation of forecasts and actuals in the evaluation sample.

3 Empirical design

This section describes the forecasting system, the pseudo out-of-sample design, and the diagnostics used to compare nested information sets under hierarchical shrinkage.

3.1 Forecasting system and notation

For each information set, I estimate the same reduced-form VAR and only change the information set.[1] Let y_t denote the $n \times 1$ vector of endogenous variables observed at time t. For the Small information set, y_t stacks the macro variables (INDPRO, CPIAUCSL, UNRATE, FEDFUNDS). For the Medium set, y_t augments the macro block with financial prices (GS10, SP500, DCOILWTICO). For the Full set, y_t further adds UMCSENT. The reduced-form VAR of order p is

$$y_t = c + \sum_{\ell=1}^{p} B_\ell y_{t-\ell} + u_t, \qquad u_t \sim \mathcal{N}(0, \Sigma), \tag{1}$$

where c is an $n \times 1$ intercept, B_ℓ are $n \times n$ coefficient matrices, and Σ is the reduced-form covariance matrix. In the empirical implementation, $p = 12$ to match the monthly data frequency and the code-based lag order.

Stacking observations over $t = 1, \ldots, T$, define the regressor matrix X with an intercept and lagged y_t terms and the coefficient matrix $\Phi = [c, B_1, \ldots, B_p]'$. The reduced-form system can be written as

$$Y = X\Phi + U, \qquad \text{vec}(U) \sim \mathcal{N}(0, I_T \otimes \Sigma), \tag{2}$$

which implies a Gaussian likelihood for (Φ, Σ). This formulation makes explicit that all equations are jointly estimated with a common covariance matrix, and it provides

[1] All quantitative evidence is generated by the hierarchical BVAR system described in this section; replication materials are summarized in the appendix.

the likelihood kernel used for the Bayesian posterior.

3.2 Hierarchical Minnesota prior

The Bayesian VAR is regularized with a Minnesota prior that shrinks the system toward a parsimonious univariate benchmark. For persistent level variables, the prior mean on the first own lag is set to one and all other coefficients are centered at zero; for stationary variables, all lag coefficients are centered at zero. The prior on the coefficient matrix takes the standard conjugate form

$$\mathrm{vec}(\Phi) \mid \Sigma, \lambda \sim \mathcal{N}(\mathrm{vec}(\Phi_0),\ \Sigma \otimes \Omega(\lambda))\,, \tag{3}$$

where Φ_0 encodes the prior means and $\Omega(\lambda)$ is a diagonal tightness matrix. For the coefficient on variable j at lag ℓ in equation i, the Minnesota prior variance is

$$\Omega_{ij,\ell}(\lambda) = \left(\frac{\lambda^2}{\ell^{2\alpha}}\right)\left(\frac{\sigma_i^2}{\sigma_j^2}\right)\psi_{ij}, \tag{4}$$

with σ_i^2 denoting a scale estimate for equation i, α governing lag decay, and ψ_{ij} applying additional shrinkage to cross-variable lags (normalized to one for own lags in the standard Minnesota design). This structure delivers stronger shrinkage on long lags and cross-variable effects while preserving flexibility for own-lag persistence.

Two additional priors stabilize persistence and initialization. A sum-of-coefficients prior favors near-unit-root dynamics in levels, and a dummy-initial-observation prior anchors the system to initial conditions (Bańbura et al., 2010; Giannone et al., 2015; Kuschnig & Vashold, 2021). These components are implemented alongside the Minnesota prior and are treated as part of the regularization mechanism rather than structural restrictions.

Hierarchical shrinkage is learned rather than tuned. The global tightness λ and lag-decay α are assigned Gamma hyperpriors with bounded support, consistent with the BVAR implementation. For example,

$$\lambda \sim \mathrm{Gamma}(a_\lambda, b_\lambda)\ \mathbb{I}(\lambda_{\min} \le \lambda \le \lambda_{\max}), \tag{5}$$

with analogous structure for α. The hyperparameters are updated via Metropolis–Hastings steps within the posterior sampler, so shrinkage adapts to the data rather than being fixed ex ante.

3.3 Marginal likelihood and hierarchical selection

The hierarchical selection logic follows the marginal likelihood principle. Conditional on λ, the likelihood integrates out Φ and Σ under the conjugate prior to yield the marginal data density $p(Y \mid \lambda)$. The posterior for λ is proportional to this marginal density times the hyperprior, and the MCMC draws concentrate on values that balance fit and complexity. This data-driven criterion penalizes over-parameterization when the information set expands and thereby stabilizes out-of-sample performance in nested comparisons.

3.4 Pseudo out-of-sample evaluation

I evaluate performance in a recursive pseudo out-of-sample design with expanding estimation windows. The initial estimation window begins in 1985M1 and ends in 2000M12, so the first forecast origin is 2001M1. Forecast origins proceed monthly through 2019M11, which yields evaluation targets through 2019M12 for the one-step horizon; RMSFEs at longer horizons are computed over the available non-missing targets implied by this alignment. At each origin, the system is re-estimated using all data available up to that origin and then produces point forecasts at the horizons reported in the main accuracy table. This recursion mirrors a real-time workflow while remaining descriptive because it uses revised data rather than real-time vintages.

3.5 Forecast accuracy and nested-model inference

Forecast accuracy is summarized by RMSFE on the common evaluation scale described in Section 2. For target i and horizon h,

$$\mathrm{RMSFE}_{i,h} = \left(P^{-1} \sum_{t \in \mathcal{T}} (y_{i,t+h} - \hat{y}_{i,t+h|t})^2 \right)^{1/2}.$$

I report RMSFEs (Table 1) and relative RMSFEs versus a random-walk (no-change) benchmark on the evaluation scale (Figure 1). Because the information sets are nested, standard equal-accuracy tests can have nonstandard behavior (Clark & McCracken, 2001); I therefore emphasize magnitudes and stability and report a nested-model-robust adjustment as a robustness check (Clark & West, 2007) (Appendix Table 4).

3.6 Forecast discipline: revision-based diagnostic

To assess whether forecast updates are systematically related to subsequent forecast errors, I use the error-on-revision regression framework of Coibion and Gorodnichenko (2015) applied to model-implied forecasts:

$$(z_{t,h} - \hat{z}^{(m)}_{t,h|t}) = \alpha_h + \beta_h r^{(m)}_{t,h} + \varepsilon_{t,h}, \tag{6}$$

where $r^{(m)}_{t,h}$ is the revision to the forecast for the same target date made one period apart. In this paper, the regression is used as a diagnostic of the forecasting system's updating rule: it measures whether revisions are followed by predictable errors, indicating systematic patterns in updating. Because the forecasts are produced under shrinkage, such patterns can partly reflect prior-induced conservatism, misspecification, or instability rather than an economic mechanism.

4 Results

This section synthesizes evidence from forecast accuracy and revision diagnostics, then interprets the patterns and limitations within a signal-extraction framework. The accuracy results suggest that adding financial prices can be associated with smaller errors than a macro-only information set, while the step from Medium to Full yields limited and unstable incremental evidence for sentiment (Table 1; Figure 1). The revision diagnostic indicates systematic updating patterns that vary with the information set even when point accuracy changes little (Table 2; Figure 2). Because the specifications are nested, these comparisons are read as descriptive magnitudes, with nested-robust checks reported in the appendix (Table 4).

Two interpretive cautions guide the discussion. First, nested-model comparisons are susceptible to overfitting and nonstandard test behavior, so the emphasis is on the direction and stability of patterns rather than sharp hypothesis testing. Second, the evaluation uses revised data rather than real-time vintages, so the exercise is best viewed as a disciplined comparison of information sets rather than a literal replication of real-time forecasting performance.

4.1 Forecast accuracy

Table 1 reports RMSFEs by information set, and Figure 1 summarizes the same comparison in relative terms versus a random-walk (no-change) benchmark on the evaluation scale. The comparison between Medium and Full isolates sentiment's incremental contribution conditional on prices, while the comparison between Small and Medium captures the incremental content of financial prices. The evidence provides limited and unstable incremental support for sentiment once prices are included, consistent with information overlap across forward-looking indicators.

Viewed through a signal-to-noise lens, price-based variables can embed a cleaner signal about latent fundamentals than survey sentiment. Conditional on prices, the remaining independent signal in sentiment can be weak relative to measurement noise, so the marginal gain in RMSFE is small even when sentiment is correlated with the target. In a hierarchical BVAR, this redundancy shows up as posterior shrinkage of sentiment coefficients toward zero when the data do not support additional explanatory power.

From a forecasting-practice perspective, adding redundant predictors increases dimensionality without improving accuracy, raising estimation variance and requiring stronger regularization to maintain stability. The hierarchical prior mitigates this cost by tightening as the information set expands, but it cannot create signal where the data indicate none.

A natural interpretation of the Small-to-Medium comparison is information aggregation: financial prices may embed forward-looking signals that are not fully captured by macro aggregates alone. Where the addition of prices improves accuracy, it is consistent with price discovery adding marginal predictive content in a linear forecasting system, though the magnitude and stability of the gains are heterogeneous across targets and horizons.

The Medium-to-Full comparison is a signal-extraction exercise. If sentiment surveys are noisy or collinear with price-based signals, their marginal contribution to point-forecast accuracy can be small even when sentiment contains information about future fundamentals. This framing emphasizes overlap and measurement noise rather than a claim that sentiment is uninformative in general.

Table 1: Root Mean Squared Forecast Errors

model	variable	h1	h3	h12
Small	CPI	3.468	2.643	1.305
Small	INDPRO	7.649	5.558	4.998
Medium	CPI	2.982	2.500	1.349
Medium	INDPRO	7.315	4.966	4.371
Full	CPI	3.128	2.538	1.330
Full	INDPRO	7.424	5.087	4.387

Notes: RMSFEs are computed from recursive pseudo out-of-sample forecasts on the common evaluation scale described in Section 2. Information sets are nested, so interpretation emphasizes incremental information content rather than structural effects. .

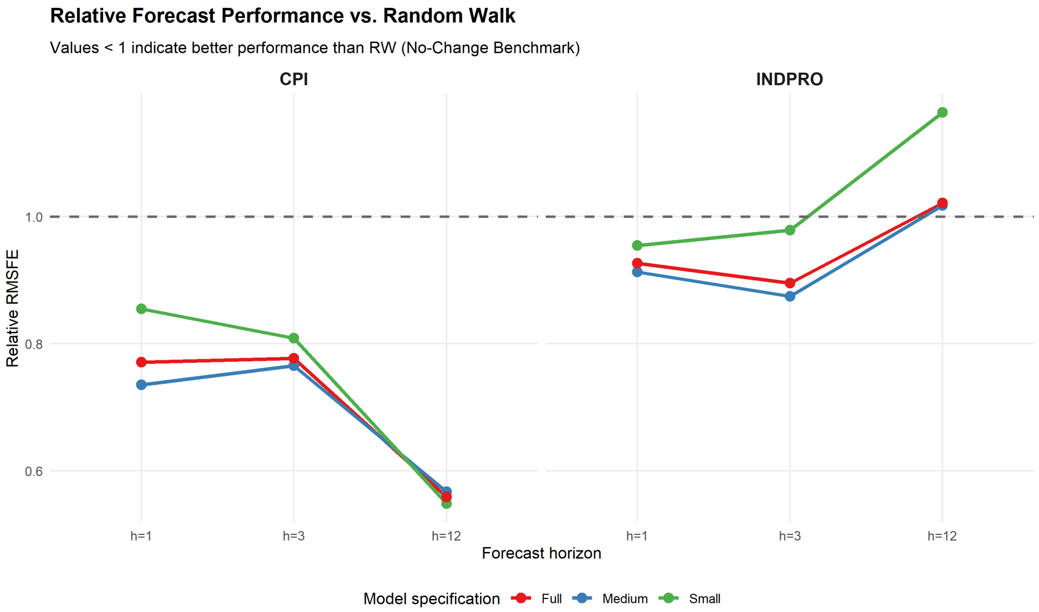


Figure 1: Relative forecast accuracy versus a random-walk (no-change) benchmark

Notes: Relative RMSFEs versus a random-walk benchmark in growth rates (no-change forecast equals zero on the evaluation scale). .

4.2 Forecast discipline: revision-based diagnostic

Table 2 reports error-on-revision coefficients from the Coibion and Gorodnichenko (2015) diagnostic applied to model-implied forecasts, and Figure 2 visualizes the same patterns. For inflation, the coefficients change sign across horizons; for real activity, estimates are imprecise and do not show a stable revision–error association. Interpreted within this model-based diagnostic, $\beta_h = 0$ corresponds to an efficient updating benchmark, while statistically detectable deviations indicate that revisions are predictably related to subsequent errors, consistent with underreaction when

$\beta_h > 0$ and overreaction when $\beta_h < 0$. This is evidence about the forecasting system's internal use of its own history rather than a structural statement about households or firms.

Interpreted as systematic updating behavior, a positive β_h indicates conservative updating inertia: revisions incorporate new information only partially, so errors are predictable in the direction of the revision. In a hierarchical BVAR, tighter overall shrinkage pulls coefficients toward persistence and can induce such inertia, especially when the data provide limited independent signal. Conversely, negative β_h can arise if the system overreacts to transient signals in the information set. The variation in signs and magnitudes across information sets is therefore consistent with the interaction between data-driven shrinkage and the informational redundancy of the predictors.

The diagnostic is intentionally narrow. It is informative about the internal consistency of the updating rule, but it does not identify the sources of any predictability in revisions. Because forecasts are produced under hierarchical shrinkage, the revision–error association can reflect conservative updating, model misspecification, or shifting data relationships. The analysis therefore emphasizes whether the patterns move with the information set rather than attributing them to a particular behavioral channel.

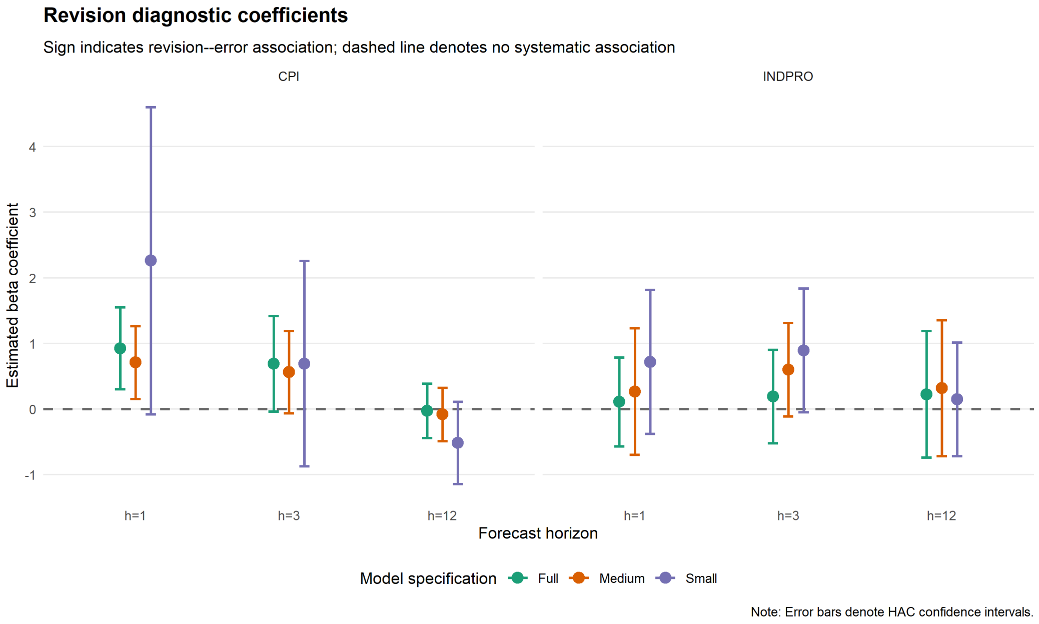


Figure 2: Revision diagnostic coefficients across information sets

Notes: Revision diagnostic coefficients from model-implied forecasts; interpreted as internal updating patterns in the regularized system. .

Table 2: Coibion–Gorodnichenko Regression Results

term	estimate	std.error	statistic	p.value
Small CPI h=1	2.2608	1.1942	1.8931	0.0597
Small CPI h=3	0.6917	0.7987	0.8661	0.3874
Small CPI h=12	-0.5178	0.3213	-1.6115	0.1086
Small INDPRO h=1	0.7184	0.5612	1.2801	0.2019
Small INDPRO h=3	0.8923	0.4816	1.8528	0.0653
Small INDPRO h=12	0.1449	0.4423	0.3276	0.7436
Medium CPI h=1	0.7086	0.2840	2.4951	0.0134
Medium CPI h=3	0.5602	0.3204	1.7485	0.0818
Medium CPI h=12	-0.0841	0.2065	-0.4073	0.6842
Medium INDPRO h=1	0.2663	0.4916	0.5416	0.5886
Medium INDPRO h=3	0.5983	0.3637	1.6449	0.1015
Medium INDPRO h=12	0.3184	0.5292	0.6017	0.5480
Full CPI h=1	0.9257	0.3188	2.9040	0.0041
Full CPI h=3	0.6894	0.3729	1.8488	0.0659
Full CPI h=12	-0.0272	0.2111	-0.1291	0.8974
Full INDPRO h=1	0.1078	0.3465	0.3110	0.7561
Full INDPRO h=3	0.1889	0.3630	0.5204	0.6033
Full INDPRO h=12	0.2237	0.4925	0.4543	0.6501

Notes: Error-on-revision regression following Coibion and Gorodnichenko (2015) applied to model-implied forecasts. Within this diagnostic, $\beta_h = 0$ is the efficient-updating benchmark; departures from zero imply underreaction or overreaction in the model's updating rule rather than structural evidence on information rigidity. .

4.3 Regularization and model stability

Figure 3 reports the evolution of the posterior mean of the hierarchical shrinkage tightness parameter. The key message is methodological: hierarchical regularization adapts the forecasting system's effective complexity as information sets expand and as the data environment changes, which helps make horse-race comparisons less sensitive to ad hoc tuning choices.

The tightness parameter is a statistical object governing the strength of prior shrinkage. Changes in its posterior mean reflect how strongly the data support deviations from the prior, not shifts in economic behavior or beliefs. This interpretation keeps the role of regularization anchored in statistical discipline.

By letting the prior adapt to information set size and data fit, hierarchical shrinkage improves comparability across models. It limits the risk that larger information sets appear to perform well simply because they overfit in-sample variation, which is especially important in nested comparisons.

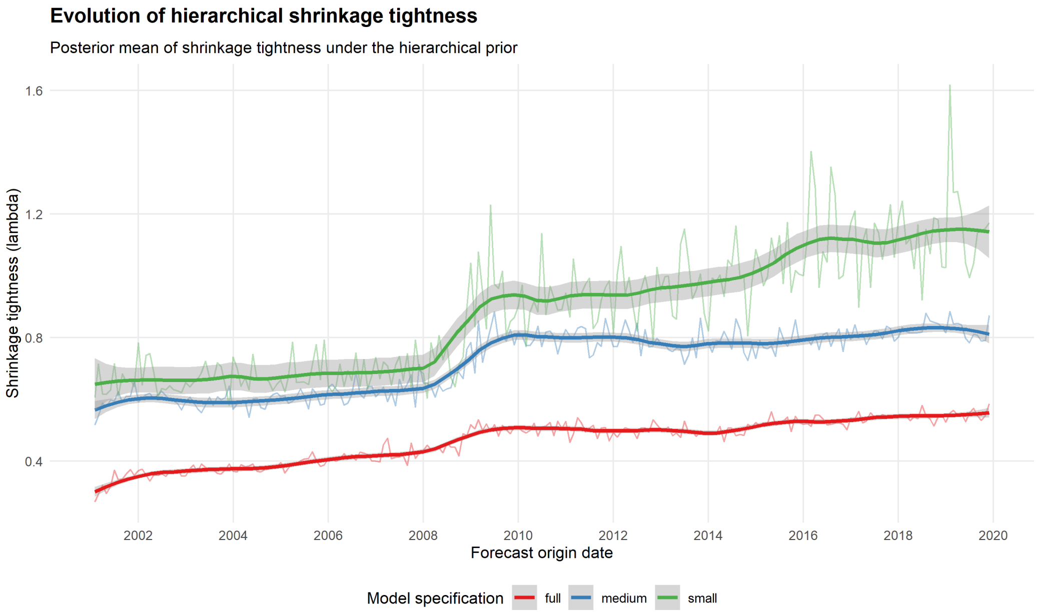


Figure 3: Evolution of hierarchical shrinkage tightness

Notes: Posterior mean of hierarchical shrinkage tightness over recursive forecast origins; reflects statistical regularization rather than behavioral change. .

4.4 Economic interpretation and mechanisms

The evidence that sentiment offers limited incremental accuracy aligns with a signal-extraction view of information aggregation. If financial prices efficiently aggregate dispersed information, they can act as a sufficient statistic for forward-looking fundamentals within a linear forecasting system. By contrast, survey sentiment reflects dispersed signals filtered through household information-processing costs and measurement error. Conditional on prices, the marginal signal in sentiment is therefore difficult to distinguish from noise in point-forecast performance.

A compact signal-extraction sketch clarifies the intuition. Let the macro target be driven by a latent state and noise, and let prices and sentiment be noisy measurements of that state:

$$y_t = s_t + \varepsilon_t, \qquad p_t = s_t + \nu_t, \qquad m_t = s_t + \eta_t.$$

The relevant object is the signal-to-noise ratio (SNR) of each observable. When the SNR of price-based indicators is higher than the SNR of sentiment, conditioning on p_t absorbs most of the latent variation in s_t that is forecast-relevant. In that case, the optimal linear weight on m_t conditional on p_t is small, and the remaining contribution of sentiment is largely noise.

This logic maps directly into the hierarchical BVAR. When the data indicate that m_t adds little incremental explanatory power once p_t is included, the posterior distribution of sentiment coefficients concentrates near zero, reinforced by the Minnesota prior's cross-variable shrinkage. The resulting forecasts can therefore show limited RMSFE gains from sentiment even if sentiment is correlated with the target, because the correlation is already captured by prices in the information set.

The signal-extraction interpretation is limited to the information set and linear specification used here. Alternative financial variables, nonlinear dynamics, or distributional forecasting objectives could alter the incremental value of sentiment, so the findings should be read as model- and sample-specific. Consistent with this perspective, revision diagnostics may still move with the information set even when average point accuracy changes little, because revisions reflect the system's internal updating rule rather than marginal predictive content alone.

Long-horizon inflation forecasts are dominated by slow-moving trends and persistence, making parsimonious benchmarks competitive. Hierarchical shrinkage encourages persistence and can dampen short-run updates, so revision diagnostics may display systematic updating patterns that reflect regularization, misspecification, or structural change rather than beliefs. This mechanism view is interpretive rather than identified: the regression is a diagnostic of the forecasting system, and the evidence is consistent with sentiment affecting the pattern of revisions more than average point accuracy.

4.5 Limitations

The revision regression is a diagnostic tool, and its patterns can arise from shrinkage, misspecification, or structural change. The nested design limits power for incremental comparisons, and alternative sentiment measures, real-time data vintages, or density-forecast evaluation could alter the conclusions. Overall, the evidence is consistent with limited and unstable incremental support for sentiment in point-forecast accuracy and with revision diagnostics that vary with the information set, but the comparisons remain descriptive.

A further limitation is that evaluation is performed on revised data. Real-time data availability and revisions can affect both forecast accuracy and revision diagnostics, so the results should not be interpreted as a definitive assessment of real-time operational performance.

5 Conclusion

This paper evaluates whether consumer sentiment adds incremental predictive content in a hierarchical BVAR with nested information sets and a revision-based diagnostic.

The central conclusion is a hierarchy for linear point forecasting: financial prices capture most forward-looking information beyond macro aggregates, while sentiment delivers, at most, marginal and unstable incremental gains once prices are included. This pattern is consistent with a signal-extraction view and with posterior shrinkage of redundant predictors, and it persists under nested-robust checks in Appendix Table 4, which reinforce the need for cautious interpretation rather than overstatement of small differences.

The revision diagnostic provides complementary evidence on internal updating behavior. When β_h departs from zero, the system's revisions are predictably related to subsequent errors, indicating conservative updating or overreaction relative to an efficient-updating benchmark within the model. This remains a statement about the forecasting rule rather than about household or firm behavior.

From a practical perspective, the results suggest that sentiment surveys may be more informative as contextual indicators or for monitoring forecast revisions than as incremental predictors in price-augmented linear systems. This implication is framed as forecasting guidance, not a statement about the welfare relevance of sentiment.

Future work can assess state dependence, alternative sentiment measures, real-time data vintages, and density-forecast evaluation. These extensions would clarify whether the limited incremental role of sentiment for point forecasts is robust to different data environments and forecasting objectives.

References

Atkeson, A., & Ohanian, L. E. (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, *25*(1), 2–11.

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector autoregressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2018). Diagnostic expectations and credit cycles. *American Economic Review*.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Diagnostic expectations and the macroeconomy. *Journal of Economic Perspectives*.

Bram, J., & Ludvigson, S. (1998). Does consumer confidence forecast household expenditure? a sentiment index horse race. *Federal Reserve Bank of New York Economic Policy Review*, *4*(2), 59–78.

Carroll, C. D., Fuhrer, J. C., & Wilcox, D. W. (1994). Does consumer sentiment forecast household spending? if so, why? *American Economic Review*, *84*(5), 1397–1408.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110. doi: 10.1016/S0304-4076(01)00083-9

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311. doi: 10.1016/j.jeconom.2006.05.023

Coibion, O., & Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, *105*(8), 2644–2678. doi: 10.1257/aer.20110306

Croushore, D. (2005). Do consumer confidence indexes help forecast consumer spending in real time? *Journal of Business & Economic Statistics*.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. doi: 10.1162/REST_a_00483

Kuschnig, N., & Vashold, L. (2021). Bvar: Bayesian vector autoregressions with hierarchical prior selection in r. *Journal of Statistical Software*, *100*(14), 1–27. doi: 10.18637/jss.v100.i14

Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives*, *18*(2), 29–50.

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

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(s1), 3–33.

A Data definitions

Table 3: Information sets and data definitions

Set	Variable	Source	Transformation	Frequency
Small	INDPRO	FRED	log level	Monthly
Small	CPIAUCSL	FRED	log level	Monthly
Small	UNRATE	FRED	level	Monthly
Small	FEDFUNDS	FRED	level	Monthly
Medium	GS10	FRED	level	Monthly
Medium	SP500	Yahoo Finance	log level	Monthly
Medium	DCOILWTICO	FRED	log level	Monthly
Full	UMCSENT	FRED (U. Michigan Surveys)	level	Monthly

Notes: Sample period 1985M1 - 2019M12 for all series. Transformations refer to the level used in estimation; evaluation uses a common growth-rate scale as described in Section 2. Sources: FRED (Federal Reserve Bank of St. Louis) for macro and financial series, Yahoo Finance for the S&P 500 index.

B Additional figures and robustness

Nested-model forecast accuracy: Clark–West tests. Table 4 reports Clark–West MSPE-adjusted tests for nested model comparisons (Small vs. Medium; Medium vs. Full). This robustness addresses the nonstandard behavior of standard equal-accuracy tests under nesting.

Table 4: Clark–West (2007) MSPE-Adjusted Tests for Nested Models

Smaller	Larger	variable	horizon	t-stat	p-value	N	NW lag
Small	Medium	CPI	$h = 1$	3.312***	0.001	227	1
Small	Medium	CPI	$h = 3$	2.405***	0.008	225	3
Small	Medium	CPI	$h = 12$	-0.063	0.525	216	12
Small	Medium	INDPRO	$h = 1$	3.211***	0.001	227	1
Small	Medium	INDPRO	$h = 3$	2.387***	0.009	225	3
Small	Medium	INDPRO	$h = 12$	2.452***	0.008	216	12
Medium	Full	CPI	$h = 1$	-1.146	0.874	227	1
Medium	Full	CPI	$h = 3$	-0.325	0.627	225	3
Medium	Full	CPI	$h = 12$	0.742	0.230	216	12
Medium	Full	INDPRO	$h = 1$	0.107	0.458	227	1
Medium	Full	INDPRO	$h = 3$	0.057	0.477	225	3
Medium	Full	INDPRO	$h = 12$	0.253	0.400	216	12

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Clark–West (2007) MSPE-adjusted test for equal forecast accuracy in nested models. For smaller-model forecast error $e_{1t} = y_t - f_{1t}$ and larger-model error $e_{2t} = y_t - f_{2t}$, the adjusted loss differential is $d_t = e_{1t}^2 - \left(e_{2t}^2 - (f_{2t} - f_{1t})^2\right)$. The test regresses d_t on a constant. Newey–West HAC standard errors use lag truncation equal to the forecast horizon (overlap adjustment). One-sided p-values reported for the alternative that the larger model improves MSPE.

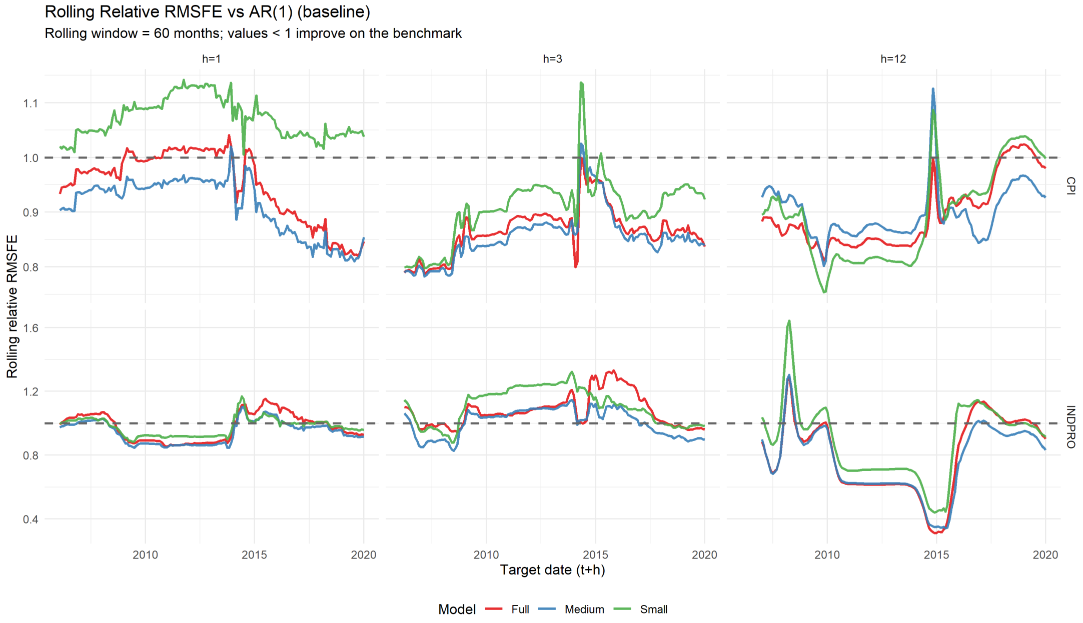


Figure 4: Rolling relative RMSFE versus an AR(1) benchmark

Notes: Rolling relative RMSFEs versus an AR(1) benchmark estimated on the evaluation-scale growth rates. .

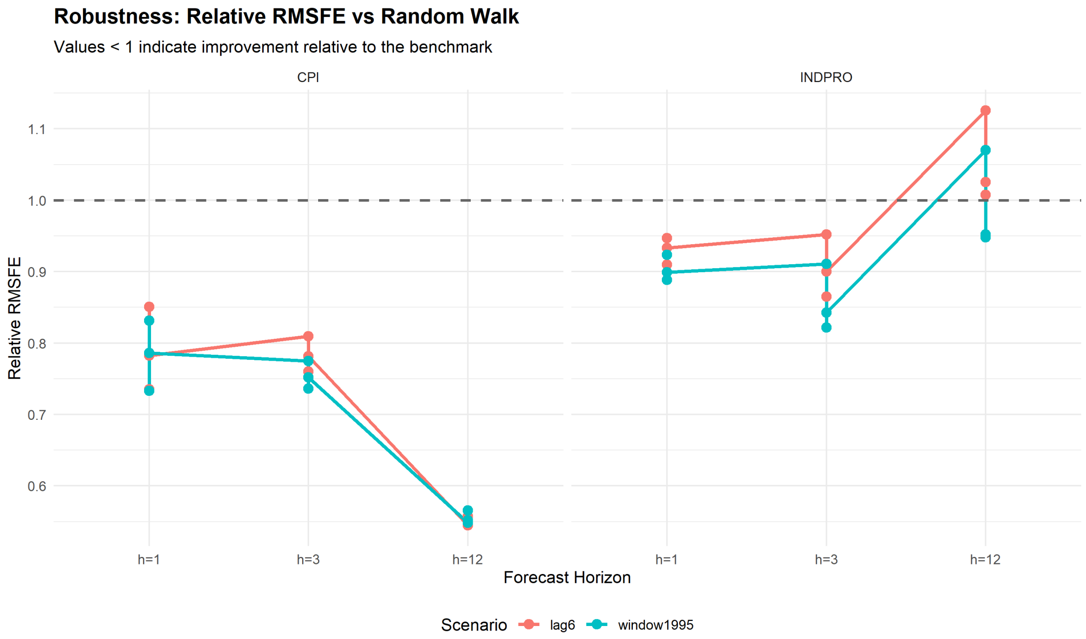


Figure 5: Robustness: relative RMSFE versus no-change benchmark

Notes: Relative RMSFEs under alternative implementation choices versus a no-change (random-walk) benchmark on the evaluation scale. .

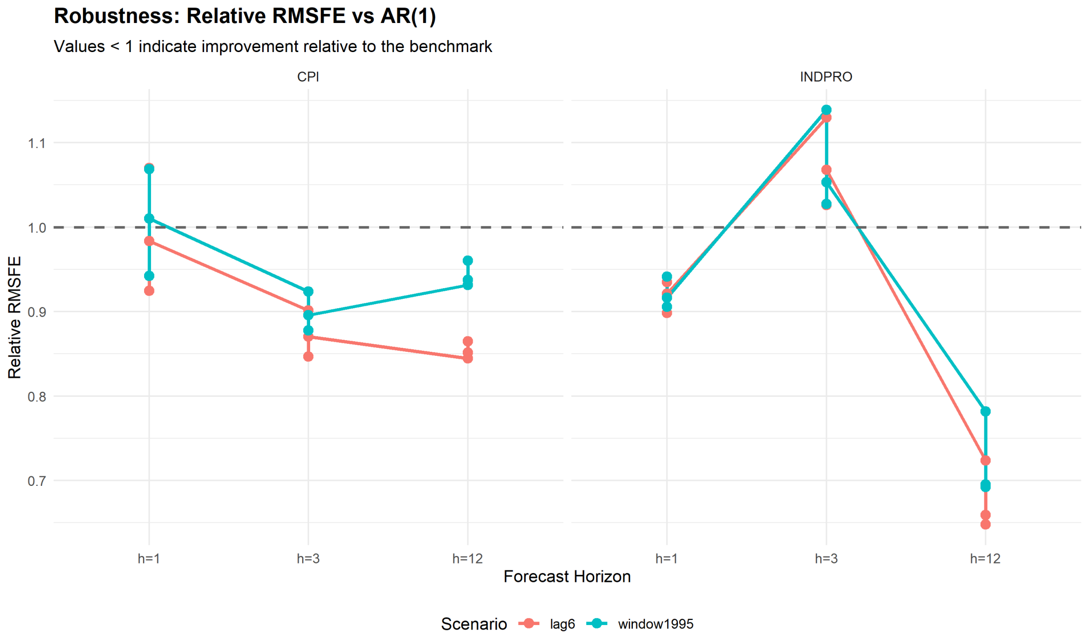
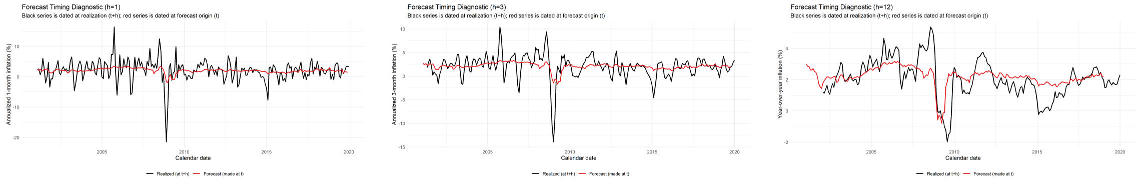


Figure 6: Robustness: relative RMSFE versus an AR(1) benchmark

Notes: Relative RMSFEs under an alternative implementation choice versus an AR(1) benchmark estimated on the evaluation-scale growth rates. .



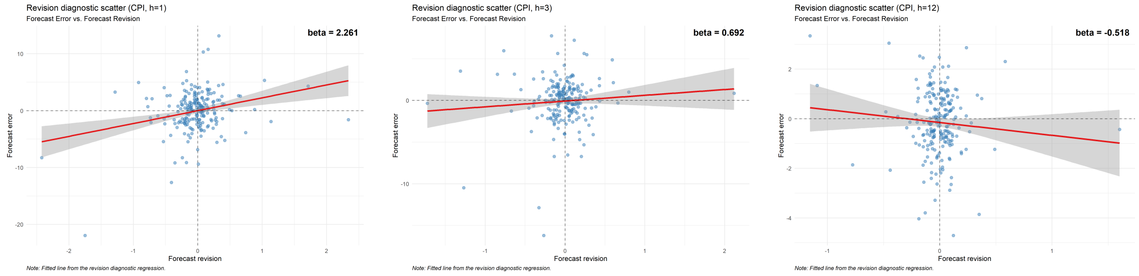
(a) Short horizon

(b) Intermediate horizon

(c) Long horizon

Figure 7: Forecast timing diagnostic (multiple horizons)

Notes: Realizations are dated at the target date and forecasts are dated at the origin date. .



(a) Short horizon

(b) Intermediate horizon

(c) Long horizon

Figure 8: Revision diagnostic scatter (multiple horizons)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation; the fitted line corresponds to the Coibion and Gorodnichenko (2015) diagnostic. .