

Geneva Graduate Institute (IHEID)

Topics in Econometrics (EI137)

Term Paper

Forecasting Horse Races and “Belief Distortions”

Hierarchical Bayesian VAR Study with Sentiment Signals

Jingle Fu

Professor: Marko Mlikota

Abstract

Does consumer sentiment improve forecasts of inflation and real activity beyond the information in macro aggregates and financial prices? I address this question using a hierarchical Bayesian VAR with nested information sets, expanding-window pseudo out-of-sample forecasting, and a revision-based diagnostic following Coibion and Gorodnichenko (2015). Point forecasts indicate marginal accuracy gains at best once financial prices are included. Revision diagnostics reveal updating patterns that vary with the information set, characterizing internal forecast discipline. The estimated coefficients and shrinkage patterns reflect conditional predictive relationships within a regularized system.

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

1 Introduction

How much does consumer sentiment improve forecasts of inflation and output after accounting for standard macro aggregates and financial market prices? I distinguish forecast accuracy from forecast discipline to separate added information from the model's updating behavior under regularization.

I compare nested information sets within a hierarchical BVAR with data-driven shrinkage. This design accommodates changes in dimensionality without ad hoc tuning (Ba  bura, Giannone, & Reichlin, 2010; Giannone, Lenza, & Primiceri, 2015; Kuschnig & Vashold, 2021). I evaluate both point-forecast accuracy and forecast discipline. The evidence offers limited support for sentiment's incremental value once financial prices are included, while revision diagnostics reveal updating patterns that vary with the information set. Given the nested structure, I emphasize relative forecasting efficiency and stability across horizons; nested-robust Clark-West adjustments appear in the appendix (Clark & McCracken, 2001; Clark & West, 2007). I implement the Coibion and Gorodnichenko (2015) regression on model-implied forecasts to diagnose the system's updating rule.

This revision diagnostic connects to work on expectations updating, including diagnostic-expectations models (Bordalo, Gennaioli, & Shleifer, 2018, 2020). Evidence on whether confidence or sentiment adds incremental forecasting information is mixed once other indicators are included, and real-time evaluations often find gains that are limited or unstable (Bram & Ludvigson, 1998; Carroll, Fuhrer, & Wilcox, 1994; Croushore, 2005; Ludvigson, 2004). The inflation-forecasting literature emphasizes that parsimonious benchmarks are difficult to beat and that relationships shift over time, which counsels caution in interpreting small differences (Atkeson & Ohanian, 2001; Stock & Watson, 2007). Hierarchical BVAR shrinkage provides a disciplined way to compare information sets of different sizes without ad hoc tuning, which is central to the design here (Ba  bura et al., 2010; Giannone et al., 2015; Kuschnig & Vashold, 2021).

A signal-extraction framework clarifies sentiment's limited predictive role. Financial prices aggregate dispersed information efficiently and may act as a sufficient statistic for future macroeconomic fundamentals. Consumer sentiment surveys, by contrast, contain idiosyncratic noise and measurement error. The analysis tests whether, conditional on efficient price discovery, the marginal signal in noisy survey measures is statistically significant for point-forecast performance.

2 Data

The dataset is monthly and spans 1985M1–2019M12, using revised data from FRED and Yahoo Finance. Variables are chosen to capture the macro state (output, inflation, unemployment), the policy stance (short rate), forward-looking market prices (term yield, equity prices, oil), and survey-based 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 Medium-to-Full comparison therefore tests whether sentiment contributes beyond information already summarized in market prices.

Figure 1 plots three core series to illustrate their co-movement prior to the BVAR analysis. Table 1 presents summary statistics for the key variables.

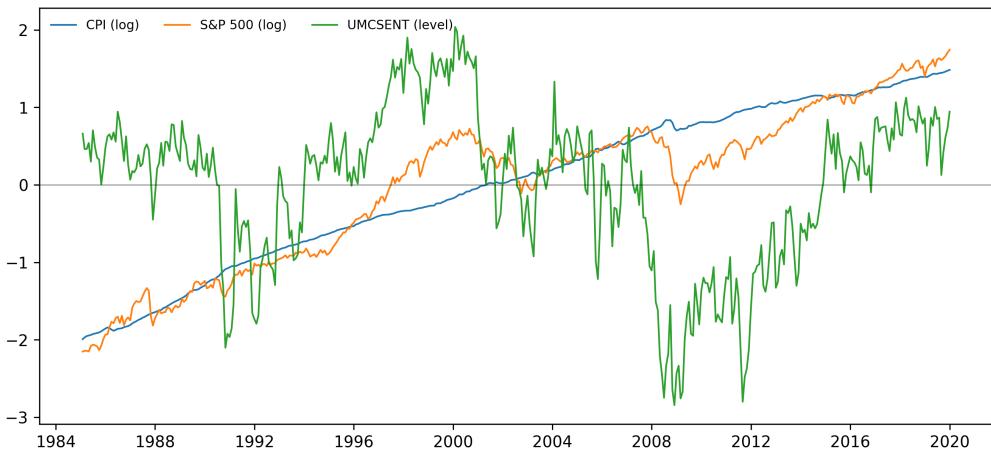


Figure 1: Key series (standardized): CPI, S&P 500, and consumer sentiment

Notes: Each series is standardized to mean zero and unit variance over the estimation sample to emphasize co-movement and regime shifts.

Table 1: Summary statistics for key variables

Variable	Mean	Std. Dev.	Corr. w/ UMCSENT
CPI (log)	5.174	0.257	-0.205
INDPRO (log)	4.412	0.209	-0.047
S&P 500 (log)	6.788	0.742	0.055
UMCSENT (level)	88.335	11.616	—

Following standard BVAR practice, the model is estimated in levels or log-levels (Giannone et al., 2015). 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,

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

yielding annualized rates of $1200(\ell_{t+1} - \ell_t)$ for $h = 1$, $400(\ell_{t+3} - \ell_t)$ for $h = 3$, and $100(\ell_{t+12} - \ell_t)$ for $h = 12$; levels are handled as cumulative changes per period.

3 Empirical design

3.1 Forecasting system and notation

I estimate the same reduced-form VAR specification for each information set, varying only the composition of y_t .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 order- p reduced-form VAR is

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

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.

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

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

implying a Gaussian likelihood for (Φ, Σ) with joint estimation of all equations.

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

3.2 Hierarchical Minnesota prior

The BVAR is regularized via 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

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

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

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

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 the system. A sum-of-coefficients prior imposes near-unit-root dynamics in levels, while a dummy-initial-observation prior anchors the system to initial conditions. These components are implemented alongside the Minnesota prior and are treated as part of the regularization mechanism that stabilizes the system.

Following Giannone–Lenza–Primiceri, I treat λ as a hyperparameter learned from the data through the marginal data density (MDD),

$$p(Y \mid \lambda) = \int p(Y \mid \Phi, \Sigma) p(\Phi, \Sigma \mid \lambda) d\Phi d\Sigma,$$

and use the BVAR package implementation to sample λ (and other prior hyperparameters) via MH-within-Gibbs steps. Conjugacy implies that, conditional on λ , the posterior is Normal–Inverse-Wishart, which yields closed-form posterior moments and a tractable posterior predictive distribution for y_{t+h} .

We first give λ a Gamma hyperprior,

$$\lambda \sim \text{Gamma}(a_\lambda, b_\lambda) \mathbb{I}(\lambda_{\min} \leq \lambda \leq \lambda_{\max}), \quad (5)$$

and, conditional on data Y , the sampler targets $p(\lambda \mid Y) \propto p(Y \mid \lambda)p(\lambda)$ within

the MH-within-Gibbs routine. Rather than a plug-in posterior mode, forecasts and diagnostics integrate over the full set of MCMC draws. The reported λ series (e.g., Figure 4) uses posterior means at each origin, so shrinkage adapts to the data rather than being fixed ex ante.

Table 2: Implementation settings and hyperparameter values

Parameter	Symbol	Specification
MCMC draws	S	$S = 10,000$
Burn-in period	B	$B = 5,000$
Minnesota tightness	λ	$\lambda \sim \Gamma(m_\lambda, s_\lambda) \mathbb{I}[0.001 \leq \lambda \leq 2.0]$ $(m_\lambda, s_\lambda) = (0.05, 0.2)$
Lag decay	α	$\alpha \sim \Gamma(m_\alpha, s_\alpha) \mathbb{I}[1 \leq \alpha \leq 3]$ $(m_\alpha, s_\alpha) = (3.0, 0.25)$
Cross-variable shrinkage	ψ	$\psi_{ij} = \psi_{ij}^{\text{BVAR}}$ $\mu_{\text{SOC}} \sim \Gamma(m_{\text{SOC}}, s_{\text{SOC}}) \mathbb{I}[0.01 \leq \mu_{\text{SOC}} \leq 50]$
Sum-of-coefficients prior	μ_{SOC}	$(m_{\text{SOC}}, s_{\text{SOC}}) = (1, 1)$ $\mu_{\text{DIO}} \sim \Gamma(m_{\text{DIO}}, s_{\text{DIO}}) \mathbb{I}[0.01 \leq \mu_{\text{DIO}} \leq 50]$
Dummy-initial-observation prior	μ_{DIO}	$(m_{\text{DIO}}, s_{\text{DIO}}) = (1, 1)$
Hyperparameter sampler		MH-within-Gibbs

3.3 Convergence and diagnostics

I assess convergence through visual inspection of hyperparameter trace behavior and by verifying stability of posterior mean shrinkage across origins, as summarized in Figure 4. The hierarchical prior mitigates overfitting as dimensionality changes, and the evidence is organized around trace stability and the evolution of posterior mean shrinkage; formal effective-sample-size or Geweke diagnostics are reserved for supplementary robustness work.

3.4 Pseudo out-of-sample evaluation

I evaluate performance using expanding-window (recursive) pseudo out-of-sample forecasts. The initial window runs from 1985M1 to 2000M12, so the first forecast origin is 2001M1. This split preserves a long pre-2001 estimation window that includes the 1990s expansion and stabilizes lag dynamics and shrinkage. The evaluation period then spans the 2001 downturn, the 2008–09 crisis, and the post-crisis low-rate environment, which are the policy-relevant regimes for forecasting comparison. Origins advance monthly through 2019M11, yielding one-step evaluation targets through 2019M12; longer-horizon RMSFEs are computed over the available non-

missing targets implied by this alignment. At each origin, the system is re-estimated with data available up to that date and then produces point forecasts at the horizons reported in the main accuracy table. This recursion mirrors a real-time workflow and evaluates performance using revised data, offering evidence on forecasting efficiency under final-vintage information.

3.5 Forecast accuracy and nested-model inference

I summarize forecast accuracy by RMSFE on the common evaluation scale described in Section 2. For target i and horizon h ,

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

Given nesting, I emphasize magnitudes and stability and report a nested-model-robust Clark–West adjustment (Appendix Table 6).

3.6 Forecast discipline: revision-based diagnostic

I assess the systematic relationship between forecast updates and subsequent errors using the error-on-revision regression of Coibion and Gorodnichenko (2015) applied to model-implied forecasts:

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

where $r_{t,h}^{(m)}$ is the revision to the forecast for the same target date made one period apart. In this paper, the regression serves 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 reflect prior-induced conservatism, misspecification, or instability as components of the updating mechanism.

4 Results

4.1 Forecast accuracy

Table 3 presents RMSFEs by information set, and Figure 2 summarizes the comparison in relative terms versus an AR(1) benchmark on the evaluation scale. The Medium-to-Full comparison isolates sentiment’s incremental value conditional on

financial prices, while the Small-to-Medium comparison captures the incremental content of financial prices. Adding sentiment to a price-augmented model yields statistically insignificant gains that are unstable across horizons. This is consistent with substantial information overlap between sentiment and forward-looking market prices.

Price-based indicators convey cleaner signals about latent fundamentals than survey sentiment. Once prices are included, the remaining independent signal in sentiment is small relative to measurement noise, which helps explain why RMSFE gains are modest even when sentiment correlates with the target. In the hierarchical BVAR, overlap between sentiment and prices is handled by regularization that can down-weight incremental predictors when the data provide little support for extra explanatory power; the small and unstable gains are consistent with that mechanism.

Redundant predictors increase dimensionality without improving accuracy, raise estimation variance, and necessitate stronger regularization to maintain stability. The hierarchical prior addresses this by tightening as the information set expands and by regularizing more aggressively when incremental signal is weak.

The improvement from the Small to the Medium model likely reflects information aggregation: financial prices embed forward-looking signals that macro aggregates alone do not fully capture. When adding prices improves accuracy, the pattern is consistent with price discovery adding marginal predictive content in a linear system, though the size and stability of the gains differ 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.

Table 3: Root Mean Squared Forecast Errors

model	variable	h1	h3	h12
Small	CPI	3.457	2.636	1.303
Small	INDPRO	7.646	5.545	4.984
Medium	CPI	2.987	2.505	1.355
Medium	INDPRO	7.322	4.968	4.378
Full	CPI	3.131	2.538	1.330
Full	INDPRO	7.424	5.083	4.388

Notes: RMSFEs are computed from expanding-window (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 in a regularized forecasting system.

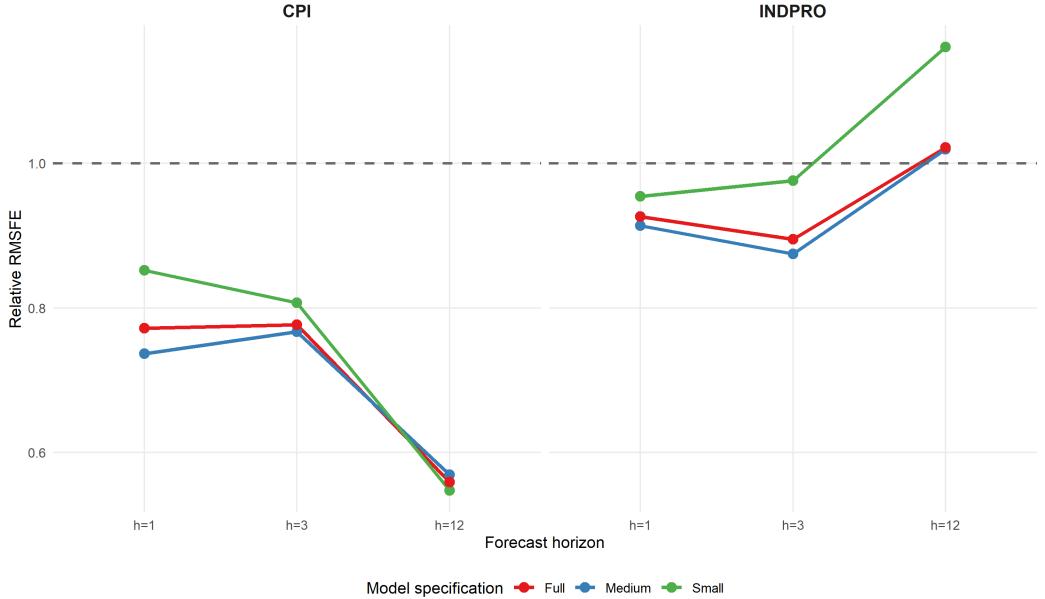


Figure 2: Relative forecast accuracy versus AR(1) benchmark

4.2 Forecast discipline: revision-based diagnostic

Table 4 presents error-on-revision coefficients from the Coibion and Gorodnichenko (2015) diagnostic applied to model-implied forecasts, and Figure 3 visualizes the same patterns. For inflation, the coefficients change sign across horizons; for real activity, estimates are imprecise and the revision–error association is unstable across horizons. Within this diagnostic, $\beta_h = 0$ serves as the 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 evidence characterizes how the forecasting system uses its own history in updating.

A positive β_h indicates conservative updating: 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 characterizes the internal consistency of the updating rule and summarizes how revisions align with subsequent errors. 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 how the patterns move with the information set and the regularization environment.

Table 4: Coibion–Gorodnichenko Regression Results

term	estimate	std.error	statistic	p.value
Small CPI h=1	2.5822	1.2343	2.0920	0.0376
Small CPI h=3	0.9300	0.8483	1.0963	0.2742
Small CPI h=12	-0.4927	0.3497	-1.4090	0.1603
Small INDPRO h=1	0.7138	0.5761	1.2390	0.2167
Small INDPRO h=3	0.9696	0.5359	1.8094	0.0718
Small INDPRO h=12	0.2366	0.5233	0.4520	0.6517
Medium CPI h=1	0.7028	0.2909	2.4157	0.0165
Medium CPI h=3	0.5494	0.3213	1.7097	0.0888
Medium CPI h=12	-0.1092	0.2117	-0.5161	0.6063
Medium INDPRO h=1	0.2198	0.4760	0.4618	0.6447
Medium INDPRO h=3	0.5940	0.3544	1.6761	0.0952
Medium INDPRO h=12	0.3340	0.5297	0.6306	0.5290
Full CPI h=1	0.9051	0.3092	2.9276	0.0038
Full CPI h=3	0.6730	0.3768	1.7863	0.0755
Full CPI h=12	-0.0661	0.2244	-0.2948	0.7684
Full INDPRO h=1	0.0772	0.3409	0.2265	0.8211
Full INDPRO h=3	0.1713	0.3461	0.4951	0.6211
Full INDPRO h=12	0.1607	0.4720	0.3404	0.7339

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 forecasting system's updating rule.

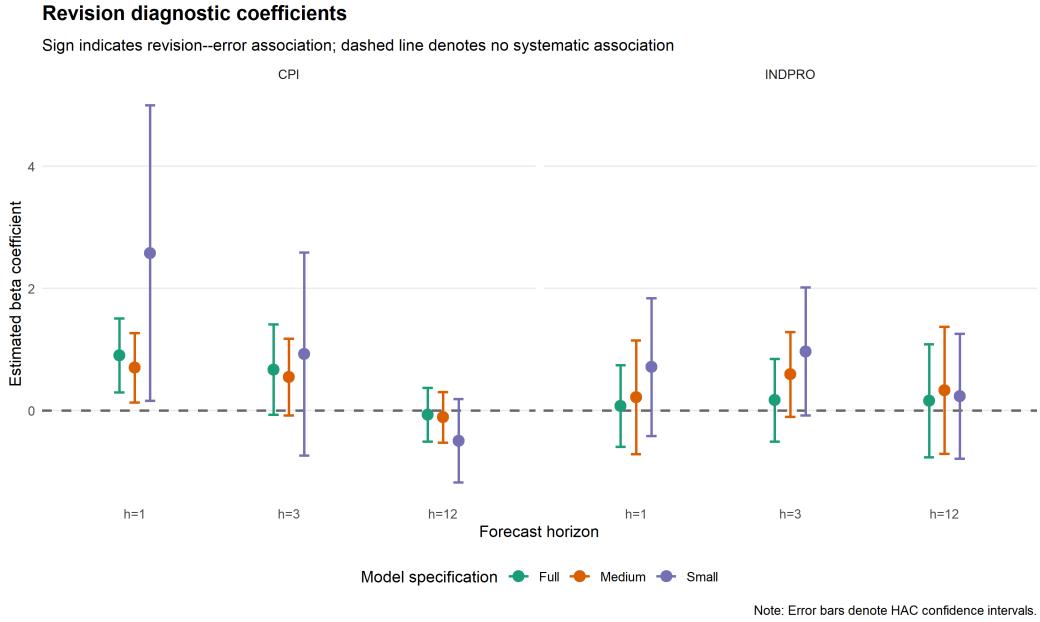


Figure 3: Revision diagnostic coefficients across information sets

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

4.3 Regularization and model stability

As shown in Figure 4 the posterior mean of the tightness parameter λ decreases monotonically from the Small to the Full model. This behavior is a direct manifestation of the signal-extraction mechanism: the hierarchical prior tightens aggressively when faced with the noisy sentiment variable, effectively down-weighting its contribution to preserve forecast efficiency. Since λ scales prior variances in (4), a lower λ implies tighter shrinkage, so the larger information set is more strongly regularized over the evaluation window. All three series increase around 2008–09 and then trend upward, suggesting moderately looser regularization over time while preserving the same cross-model ordering.

The tightness parameter governs the strength of prior shrinkage. Movements in its posterior mean indicate how much the data favor departures from the prior, keeping the interpretation of regularization grounded in the model’s statistical structure.

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.

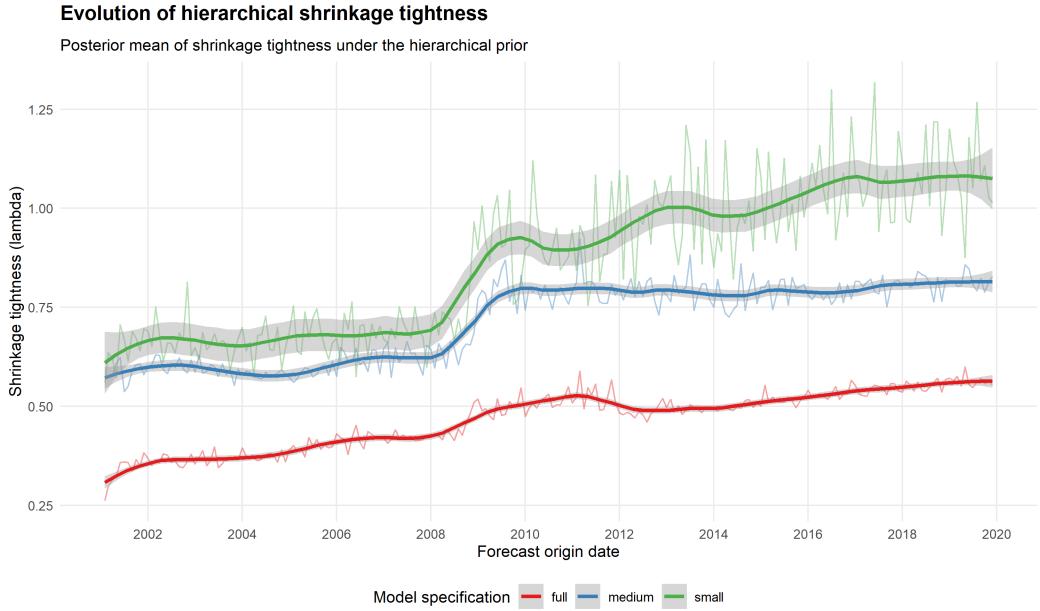


Figure 4: Evolution of hierarchical shrinkage tightness

Notes: Posterior mean of hierarchical shrinkage tightness over recursive forecast origins; reflects statistical regularization in the forecasting system.

4.4 Economic interpretation and mechanisms

Sentiment's limited contribution aligns with a signal-extraction framework: financial prices appear to aggregate dispersed information efficiently, leaving sentiment with little independent signal. 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 may contain additional noise and measurement error. Conditional on prices, the marginal signal in sentiment is therefore difficult to distinguish from noise in point-forecast performance.

Consider a latent state s_t driving the macro target y_t , with prices p_t and sentiment m_t as noisy measurements:

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

The critical quantity is each observable's signal-to-noise ratio (SNR). If price-based indicators have a higher SNR than sentiment, conditioning on p_t captures most of the forecast-relevant variation in s_t . The optimal linear weight on m_t given p_t is then small, and the residual contribution of sentiment largely reflects measurement noise.

This signal-extraction logic carries over to the hierarchical BVAR. The prior and likelihood jointly determine how much weight to place on each predictor, so when

m_t adds little incremental predictive content beyond p_t , the regularized system can assign it little effective weight. The limited and unstable RMSFE improvements from adding sentiment are consistent with that signal-extraction interpretation.

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. The regression is a diagnostic of the updating rule, and the evidence is consistent with sentiment affecting the pattern of revisions more than average point accuracy.

4.5 Limitations

Overall, the evidence indicates limited and unstable incremental support for sentiment in point-forecast accuracy and revision diagnostics that vary with the information set, framed as conditional evidence on relative forecasting efficiency.

Evaluation is performed on revised data, with real-time vintage differences left for future work. In operational forecasting, macro data arrive with publication lags and are subject to revisions, which can change the relative informational value of survey-based measures.

5 Conclusion

My central finding establishes a clear hierarchy for linear point forecasting: financial prices capture the bulk of forward-looking information beyond core macro aggregates. Consumer sentiment, by contrast, provides only marginal and unstable incremental gains conditional on financial prices. This pattern is consistent with a signal-extraction view and with regularization that down-weights redundant predictors, and it persists under nested-robust checks in Appendix Table 6, which reinforce disciplined interpretation of small differences.

Revision diagnostics complement this finding. 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 frames the evidence in terms of the forecasting rule and its updating mechanism.

These results suggest positioning sentiment surveys as contextual indicators and as inputs for monitoring forecast revisions in price-augmented linear systems, framed as forecasting guidance.

Future research should explore 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. *Journal of Finance*, 73(1), 199–227.
- Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, attention, and choice. *Quarterly Journal of Economics*, 135(3), 1399–1442.
- 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)00071-9
- 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? *The North American Journal of Economics and Finance*, 16(3), 435–450. doi: 10.1016/j.najef.2005.05.002
- 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.
- Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives*, 18(2), 29–50.
- 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 5: 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	FEDFUNDSD	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 6 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 6: Clark–West (2007) MSPE-Adjusted Tests for Nested Models

Smaller	Larger	variable	horizon	t-stat	p-value	N	NW lag
Small	Medium	CPI	$h = 1$	3.411***	0.000	227	1
Small	Medium	CPI	$h = 3$	2.343**	0.010	225	3
Small	Medium	CPI	$h = 12$	-0.175	0.569	216	12
Small	Medium	INDPRO	$h = 1$	3.223***	0.001	227	1
Small	Medium	INDPRO	$h = 3$	2.259**	0.012	225	3
Small	Medium	INDPRO	$h = 12$	2.188**	0.015	216	12
Medium	Full	CPI	$h = 1$	-1.185	0.881	227	1
Medium	Full	CPI	$h = 3$	-0.281	0.610	225	3
Medium	Full	CPI	$h = 12$	0.822	0.206	216	12
Medium	Full	INDPRO	$h = 1$	0.160	0.436	227	1
Medium	Full	INDPRO	$h = 3$	0.080	0.468	225	3
Medium	Full	INDPRO	$h = 12$	0.288	0.387	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 $d_t = e_{1t}^2 - (e_{2t}^2 - (f_{2t} - f_{1t})^2)$. 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.

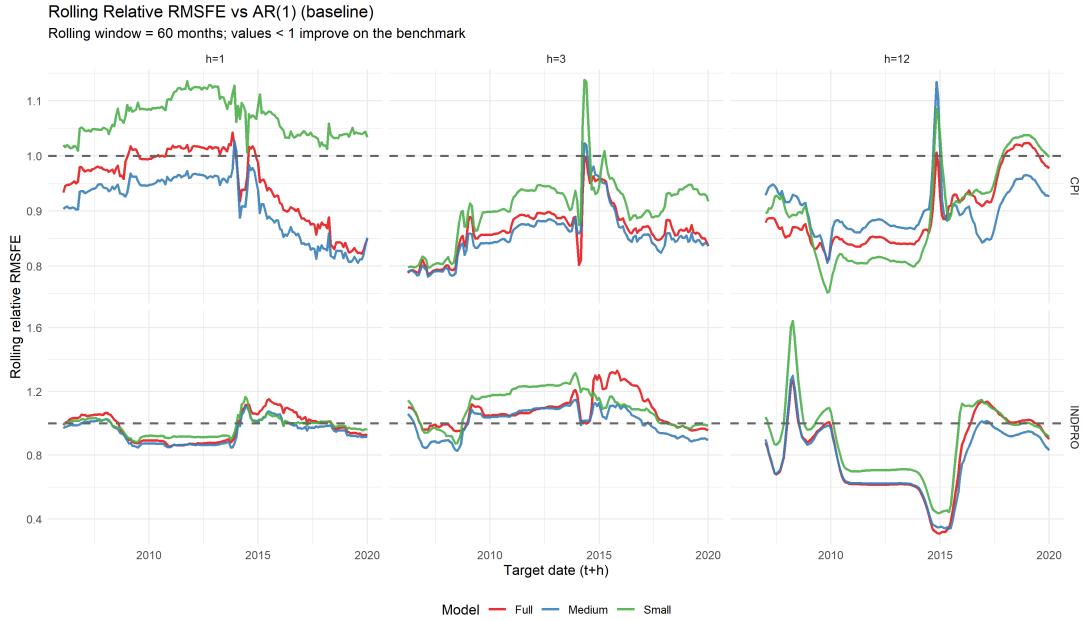


Figure 5: Rolling relative RMSFE versus an AR(1) benchmark

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

Robustness: Relative RMSFE vs Random Walk

Values < 1 indicate improvement relative to the benchmark

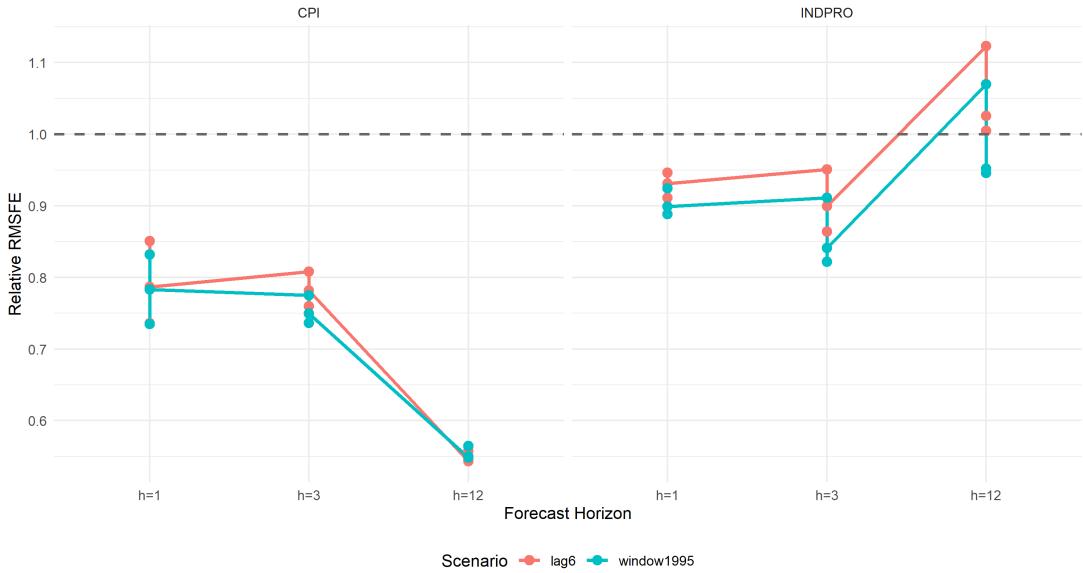


Figure 6: Robustness: relative RMSFE versus no-change benchmark

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

Robustness: Relative RMSFE vs AR(1)

Values < 1 indicate improvement relative to the benchmark

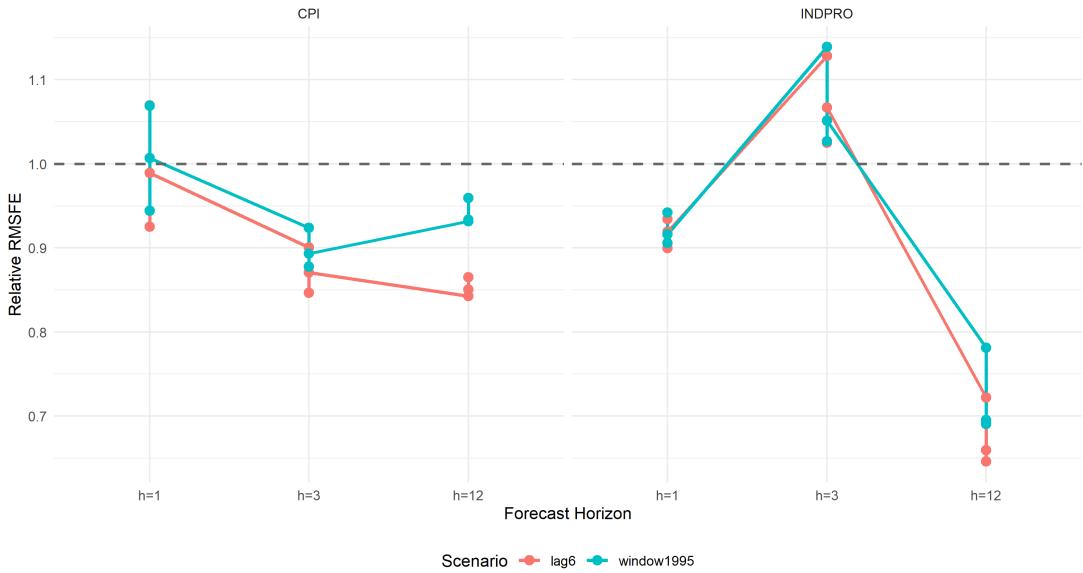
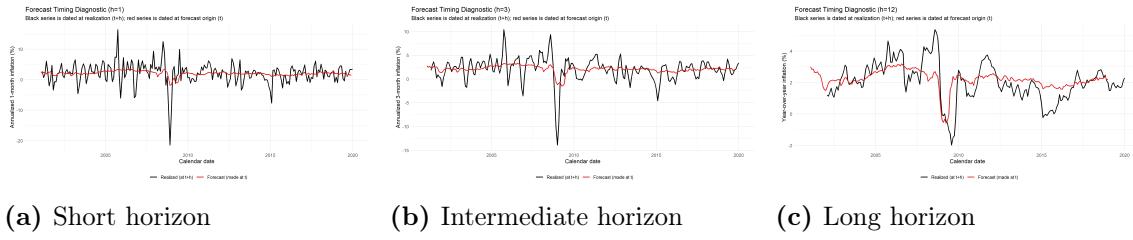


Figure 7: Robustness: relative RMSFE versus an AR(1) benchmark

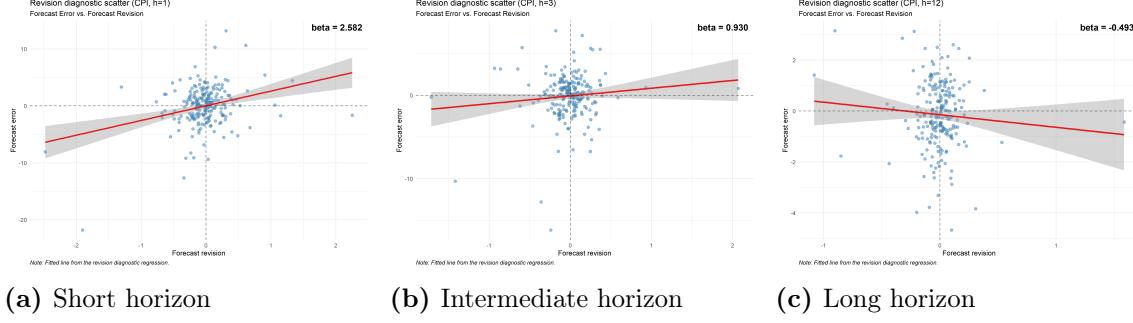
Notes: Relative RMSFEs under an alternative implementation choice versus an AR(1) benchmark estimated on the evaluation-scale growth rates.



(a) Short horizon (b) Intermediate horizon (c) Long horizon

Figure 8: 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 9: 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.