Forecasting Horse Races and "Belief Distortions": A Hierarchical Bayesian VAR Study with Sentiment Signals

Jingle Fu

December 30, 2025

Abstract

I study whether expanding the information set of a hierarchical Bayesian VAR (BVAR) to include forward-looking financial prices and consumer sentiment improves macroeconomic forecast accuracy, and whether sentiment-sensitive forecasts display systematic "belief distortions" as measured by a forecast-error-on-revision regression. The empirical design uses monthly U.S. data for 1985M1–2019M12 and an expanding-window pseudo out-of-sample exercise with forecast origins 2001M1–2019M11 at horizons $h \in \{1, 3, 12\}$. I compare three nested information sets: a small macro model, a financial-augmented model, and a full model that additionally includes consumer sentiment. The main results are strongly horizon-dependent. All BVAR specifications improve inflation forecasts relative to a no-change benchmark across horizons, while real-activity gains are concentrated at short horizons and are primarily associated with financial variables. In the revision diagnostic, the inflation coefficient declines markedly when sentiment is added at short horizons, whereas long-horizon inflation coefficients are negative in all specifications, consistent with overreaction in revisions. Model-to-model differences are estimated imprecisely in formal predictive-accuracy tests.

1. Research Question and Motivation

Macroeconomic forecasting is routinely framed as a trade-off between informational richness and statistical discipline. Augmenting the information set can in principle improve prediction by incorporating forward-looking signals, but in finite samples it can also worsen performance unless the model regularizes aggressively. Bayesian vector autoregressions (BVARs) with Minnesota-style shrinkage provide a canonical way to operationalize this trade-off in macroeconomic time series (Sims, 1980; Giannone, Lenza, and Primiceri, 2015).

This paper asks whether "soft" information in the form of consumer sentiment improves macroeconomic forecasts once one already conditions on standard macro aggregates and forward-looking financial prices, and whether sentiment-rich forecasts display systematic patterns in forecast errors that are informative about expectation formation. The second question is motivated by the forecast-error-on-revision diagnostic introduced by Coibion and Gorodnichenko (2015), which relates ex post forecast errors to forecast revisions and has been used to distinguish underreaction (information rigidity) from overreaction.

The empirical analysis is deliberately narrow. I do not propose a new estimator; instead, I provide a transparent mapping from a hierarchical BVAR implementation to out-of-sample accuracy metrics and to a revision-based diagnostic computed from the same model-generated forecasts. This mapping is valuable precisely because it forces internal consistency: the data transformations, forecasting objects, benchmarks, and inference procedures are all defined in a way that corresponds one-to-one to the project code and the produced output files.

Two hypotheses guide the empirical evaluation. *Hypothesis 1 (information set and forecast accuracy).* Expanding the information set from a small macro model to include financial prices and consumer sentiment reduces forecast loss, with the largest gains expected at short horizons for real activity (from financial prices) and at longer horizons for inflation (from sentiment). *Hypothesis 2 (sentiment and revisions).* Adding sentiment changes the coefficient linking forecast errors to forecast revisions, shifting it toward zero when sentiment reduces the systematic component of errors associated with revisions, and potentially amplifying negative coefficients at horizons where revisions overreact.

2. Data

The dataset consists of monthly U.S. time series over 1985M1–2019M12. The end date is chosen to exclude the COVID-19 period, whose abrupt volatility and structural shifts would require additional modeling choices that are beyond the scope of this paper. The series are obtained from FRED (industrial production `INDPRO`, CPI `CPIAUCSL`, unemployment `UNRATE`, federal funds rate `FEDFUNDS`, and the 10-year Treasury yield `GS10`) and from Yahoo Finance for the S&P 500 index (mapped to `SP500` in the code). Consumer sentiment is measured by the University of Michigan index `UMCSENT`.

I compare three nested information sets. The *Small* model includes `INDPRO`, `CPIAUCSL`, `UNRATE`, and `FEDFUNDS`. The *Medium* model augments the small model with `GS10` and `SP500`. The *Full* model further adds `UMCSENT`. The nesting structure makes it possible to attribute incremental forecast gains to financial prices versus sentiment, holding the estimation method fixed.

2.1 Data Transformation

In time-series analysis, (weak) stationarity is often crucial. Many macroeconomic databases (including FRED-MD) provide recommended transformations intended to remove unit roots.[1] However, the BVAR literature typically favors estimating the model in **levels** or **log-levels** (Sims, 1980; Giannone, Lenza, and Primiceri, 2015). The key reason is that Minnesota-style shrinkage can be interpreted as a structured way of regularizing persistent dynamics, including behavior close to a random walk, so that long-run comovement is not mechanically removed by differencing. If one differences the data mechanically, stationarity is ensured, but long-run equilibrium information may be attenuated.

Accordingly, we adopt the following strategy: In the estimation stage, `INDPRO`, `CPIAUCSL`, and `SP500` enter in log-levels, $x_t = \ln(X_t)$, while `UNRATE`, `FEDFUNDS`, `GS10`, and `UMCSENT` enter in levels. In the forecast-evaluation stage, point forecasts on the estimation scale are mapped into cumulative horizon-h growth rates for the log variables using the same base level at the forecast origin as in the code implementation. For horizons $h \in \{1, 3\}$ the evaluation object is annualized cumulative growth, $(1200/h)\,(x_{t+h} - x_t)$, while for $h = 12$ it is year-over-year growth, $100\,(x_{t+12} - x_t)$. This definition ensures that forecast errors compare the realized and predicted

[1] The project code follows a different convention than FRED-MD-style transformations: it estimates the BVAR in levels or log-levels and evaluates forecasts on cumulative growth rates constructed from those levels.

cumulative change from the same origin date.

For inflation based on CPIAUCSL, the evaluation target at horizon h is constructed from the log CPI level $p_t = \ln(P_t)$ as

$$\pi_{t,h} = \begin{cases} \frac{1200}{h}\left(p_{t+h} - p_t\right), & h \in \{1, 3\}, \\ 100\left(p_{t+12} - p_t\right), & h = 12, \end{cases}$$

and the same mapping is applied to industrial production growth from ln(INDPRO). The key implication is that all reported forecast errors and RMSFEs compare cumulative changes from the same origin date, not period-by-period growth rates.

2.2 Implementation with R

We employ the R package BVAR (Kuschnig and Vashold, 2021), which provides an implementation of hierarchical prior selection in the spirit of Giannone, Lenza, and Primiceri (2015).

Prior setup The prior is configured via `bv_priors()` and combines a Minnesota prior with sum-of-coefficients and dummy-initial-observation components. The overall Minnesota tightness parameter λ is treated hierarchically: rather than fixing a single tightness level, the code specifies a proper hyperprior characterized by a mode and standard deviation and then samples λ jointly with the VAR parameters in a Metropolis–Hastings step, recording posterior summaries for each forecast origin. Lag length is fixed at $p = 12$ for the baseline exercise, and distant lags are controlled primarily through shrinkage.

Recursive pseudo out-of-sample forecasting To approximate real-time forecasting, I use an expanding-window design with an initial estimation sample 1985M1–2000M12 and forecast origins running from 2001M1 through 2019M11. At each origin date, the model is re-estimated using data available up to that date, hyperparameters are re-optimized/sampled within the hierarchical framework, and multi-horizon forecasts are produced. Forecasts and auxiliary objects are saved to the output directory, including aligned forecast–actual datasets and a time series of estimated hyperparameters (`results/forecasts/hyperparameters_evolution.csv`).

3. Econometric Framework

The core methodology relies on a reduced-form VAR estimated with Minnesota-style shrinkage under a hierarchical prior. Let y_t be the vector of observables included in a given information set. For each of the three nested specifications, I estimate a BVAR with $p = 12$ monthly lags,

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

and generate multi-horizon forecasts. Forecast accuracy is assessed at horizons $h \in \{1, 3, 12\}$ using RMSFEs on the evaluation-scale targets defined in Section 2.1 and relative RMSFEs computed against two benchmarks: a no-change benchmark (random walk in levels, corresponding to a zero forecast in cumulative growth) and a univariate AR(1) benchmark estimated recursively on the same evaluation-scale growth series.

To connect the forecasting results to expectation-formation diagnostics, I also estimate the forecast-error-on-revision regression in Coibion and Gorodnichenko (2015). Let $z_{t,h}$ denote the realized evaluation-scale target (inflation or industrial production growth) from origin t to $t + h$. Let $\hat{z}^{(m)}_{t,h|t}$ denote the corresponding point forecast from model $m \in \{\text{Small}, \text{Medium}, \text{Full}\}$. Define the forecast revision as the difference between the h-step-ahead forecast made at time t and the forecast for the same target date made at $t - 1$,

$$r^{(m)}_{t,h} = \hat{z}^{(m)}_{t,h|t} - \hat{z}^{(m)}_{t,h|t-1},$$

where $\hat{z}^{(m)}_{t,h|t-1}$ is constructed as an $(h+1)$-step-ahead forecast from the previous origin so that both terms refer to $t + h$. The estimating equation is

$$(z_{t,h} - \hat{z}^{(m)}_{t,h|t}) = \alpha_h + \beta_h r^{(m)}_{t,h} + \varepsilon_{t,h}. \tag{2}$$

Because overlapping horizons induce serial correlation in residuals, inference uses Newey–West HAC standard errors with lag length equal to the forecast horizon.

3.1 Minnesota prior and hierarchical tightness

Stacking observations yields $Y = X\Phi + U$, where Φ collects $(c, B_1, \ldots, B_p)$. I impose a Minnesota-style Gaussian prior on Φ conditional on Σ:

$$\operatorname{vec}(\Phi) \mid \Sigma, \lambda \sim \mathcal{N}(\operatorname{vec}(\underline{\Phi}),\, \Sigma \otimes \underline{\Omega}(\lambda)), \qquad \Sigma \sim \mathcal{IW}(\underline{S}, \underline{\nu}),$$

where $\underline{\Phi}$ encodes the random-walk / near-random-walk belief on own first lags, and $\underline{\Omega}(\lambda)$ implements lag decay and cross-variable shrinkage. In particular, for coefficient $(B_\ell)_{ij}$,

$$\mathbb{V}\left[(B_\ell)_{ij} \mid \lambda\right] = \begin{cases} \lambda^2/\ell^2, & i = j, \\ (\lambda^2/\ell^2) \cdot (\sigma_i^2/\sigma_j^2), & i \neq j, \end{cases}$$

with σ_i^2 set from residual scales in univariate AR benchmarks.

The key departure from ad hoc calibration is that the overall tightness λ is *endogenized*. Following Giannone, Lenza, and Primiceri (2015), the code treats λ (and additional shrinkage components) as hyperparameters with proper hyperpriors and explores them via a Metropolis–Hastings step implemented in `BVAR`. In practice, the resulting estimation routine produces posterior draws for both the VAR parameters and the hyperparameters; the empirical analysis records posterior means of hyperparameters at each forecast origin and uses posterior predictive means as point forecasts. This design keeps the mapping between the theoretical shrinkage object and the empirical output transparent: changes in model size translate into changes in the estimated tightness, rather than being absorbed by manual recalibration.

3.2 Pseudo out-of-sample forecasting and evaluation

I implement an expanding-window pseudo out-of-sample exercise. The initial estimation window is 1985M1–2000M12. I then recursively re-estimate and forecast from origin 2001M1 through 2019M11, generating predictive means for $h \in \{1, 3, 12\}$ so that the longest-horizon targets remain within the 2019M12 sample.

Forecast accuracy. For target i and horizon h, compute RMSFE,

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

and report relative RMSFEs versus the no-change and AR(1) benchmarks. Differences in predictive loss are assessed using Diebold–Mariano tests (Diebold and Mariano, 1995) with Newey–West standard errors (Newey and West, 1987), following the implementation in the analysis code.

3.3 Evaluation of Forecasting Competition

We evaluate forecasting performance using the following metrics: Absolute accuracy is measured by the root mean squared forecast error (RMSFE) computed on the evaluation scale. Relative accuracy is summarized by RMSFE ratios with respect to the no-change benchmark, so that values below one indicate an improvement over a random walk in levels. Statistical comparisons are conducted using Diebold–Mariano tests applied to the squared-error loss differential, with Newey–West standard errors and lag length equal to the forecast horizon.

4. Interpretation and Expected Results

This section links the paper's hypotheses to the empirical outputs produced by the forecasting pipeline. All numerical results cited below correspond to the CSV tables in `results/tables/` and figures in `results/figures/`.

4.1 Forecast accuracy: levels of loss and relative performance

Table 1 reports RMSFEs for CPI inflation and industrial production growth across the three information sets, along with the two benchmark RMSFEs. Two patterns are immediate. First, for CPI inflation, all three BVAR specifications improve materially upon the no-change benchmark at every horizon: relative RMSFEs versus the no-change benchmark range from 0.856 to 0.873 at $h = 1$, from 0.817 to 0.821 at $h = 3$, and from 0.551 to 0.578 at $h = 12$. The full model delivers the lowest inflation RMSFE at the twelve-month horizon (1.312), consistent with sentiment containing additional longer-horizon information beyond what is already captured by macro fundamentals and financial prices.

Second, for industrial production growth the evidence favors financial prices rather than sentiment as the primary source of incremental forecasting gains at short horizons. The medium model yields the lowest RMSFE among the BVARs at $h = 1$ (7.322) and $h = 3$ (5.077). Relative to the no-change benchmark, all BVARs improve upon the benchmark at $h = 1$ and $h = 3$, but none beats the no-change benchmark at $h = 12$ (relative RMSFEs above one). Importantly, the full model still delivers the lowest $h = 12$ RMSFE among the BVARs (4.610), suggesting that sentiment can improve long-horizon industrial production forecasts relative to other multivariate specifications even when a no-change benchmark remains difficult to beat in this sample.

Table 1: Forecast accuracy across information sets

		$h = 1$	$h = 3$	$h = 12$
Panel A. RMSFE (evaluation scale)				
Small	CPI	3.482	2.671	1.341
Medium	CPI	3.474	2.683	1.375
Full	CPI	3.542	2.670	1.312
Small	INDPRO	7.633	5.536	5.084
Medium	INDPRO	7.322	5.077	4.817
Full	INDPRO	7.534	5.245	4.610
RW benchmark	CPI	4.057	3.267	2.381
AR(1) benchmark	CPI	3.226	2.933	1.535
RW benchmark	INDPRO	8.012	5.680	4.294
AR(1) benchmark	INDPRO	8.117	4.788	6.678
Panel B. Relative RMSFE vs no-change benchmark				
Small	CPI	0.858	0.818	0.563
Medium	CPI	0.856	0.821	0.578
Full	CPI	0.873	0.817	0.551
Small	INDPRO	0.953	0.975	1.184
Medium	INDPRO	0.914	0.894	1.122
Full	INDPRO	0.940	0.923	1.074

Notes: Panel A reports RMSFEs computed from the expanding-window pseudo out-of-sample forecasts. The no-change benchmark corresponds to a random walk in levels (zero forecast on the cumulative-growth evaluation scale). The AR(1) benchmark is estimated recursively on the evaluation-scale growth series for each horizon. Panel B reports RMSFEs relative to the no-change benchmark.

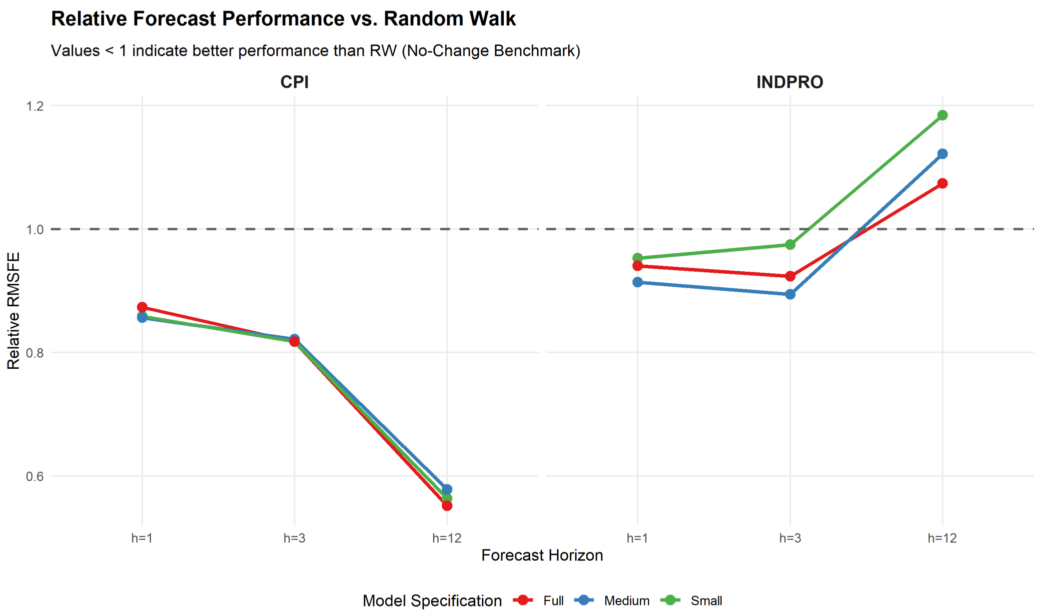

Figure 1: Relative RMSFE versus no-change benchmark (random walk)

Notes: The figure plots RMSFE for each model divided by the RMSFE of the no-change benchmark at each horizon. Values below one indicate improvement over the benchmark. The plotted values correspond to Table 1, Panel B.

4.2 Predictive-accuracy tests and what they do (and do not) show

Formal comparisons of predictive accuracy use Diebold–Mariano tests implemented on squared-error loss differentials with Newey–West standard errors. Two points are useful for interpreting the resulting p-values. First, relative to the no-change benchmark, CPI forecasts exhibit statistically meaningful improvements at the one-month horizon in all three BVAR specifications (for example, the small model yields $t = -3.12$ with $p = 0.002$). Second, relative to the AR(1) benchmark, BVAR inflation forecasts do not dominate at the one-month horizon; for the small model the DM statistic is $t = 2.47$ ($p = 0.014$), indicating significantly higher loss than the AR(1) benchmark at that horizon. Across the three BVAR specifications, DM tests comparing models pairwise generally do not reject equal predictive accuracy at conventional levels, even when point RMSFEs differ. This aligns with the broader message of the horse race: the data are informative about which information sets matter at which horizons, but not about sharp, statistically separated rankings among the multivariate models.

4.3 Revisions and "belief distortions"

Table 2 reports estimates of the forecast-error-on-revision regression. For inflation, the coefficient on revisions is large and positive at $h = 1$ for the small and medium models (2.41 and 1.92) and declines substantially once sentiment is included (0.85). In the Coibion and Gorodnichenko (2015) framework, a positive coefficient indicates that revisions are associated with forecast errors of the same sign, which is consistent with underreaction: forecasters revise in response to news but not enough, leaving ex post errors aligned with the revision. The magnitude of the coefficient therefore suggests pronounced underreaction in the smaller information sets at short horizons, while the sentiment-augmented specification attenuates this systematic association.

At the twelve-month horizon, inflation coefficients are negative in all three models (between -0.55 and -0.64), with statistical significance for the medium and full specifications. This pattern is consistent with overreaction in long-horizon revisions: upward revisions tend to be followed by negative errors, and vice versa. For industrial production, coefficients are positive but imprecisely estimated across horizons, and the evidence does not support strong conclusions about the sign patterns.

Table 2: Forecast error on forecast revision (CG regression)

Model	Target	Horizon	$\hat{\beta}_h$	SE	t	p	N
Small	CPI	$h = 1$	2.408	1.182	2.04	0.043	215
Medium	CPI	$h = 1$	1.917	0.670	2.86	0.005	215
Full	CPI	$h = 1$	0.852	0.445	1.91	0.057	215
Small	CPI	$h = 3$	0.684	0.759	0.90	0.369	215
Medium	CPI	$h = 3$	0.581	0.506	1.15	0.253	215
Full	CPI	$h = 3$	0.256	0.587	0.44	0.663	215
Small	CPI	$h = 12$	−0.559	0.300	−1.87	0.063	215
Medium	CPI	$h = 12$	−0.548	0.267	−2.05	0.041	215
Full	CPI	$h = 12$	−0.637	0.320	−1.99	0.048	215
Small	INDPRO	$h = 1$	0.767	0.595	1.29	0.199	215
Medium	INDPRO	$h = 1$	0.620	0.506	1.23	0.222	215
Full	INDPRO	$h = 1$	0.320	0.379	0.84	0.400	215
Small	INDPRO	$h = 3$	0.863	0.480	1.80	0.074	215
Medium	INDPRO	$h = 3$	0.799	0.414	1.93	0.055	215
Full	INDPRO	$h = 3$	0.274	0.412	0.66	0.507	215
Small	INDPRO	$h = 12$	0.114	0.497	0.23	0.818	215
Medium	INDPRO	$h = 12$	0.305	0.482	0.63	0.528	215
Full	INDPRO	$h = 12$	0.229	0.440	0.52	0.603	215

Notes: The dependent variable is the forecast error and the regressor is the forecast revision, both constructed from model-implied forecasts on the evaluation scale. Standard errors are Newey–West with lag h.

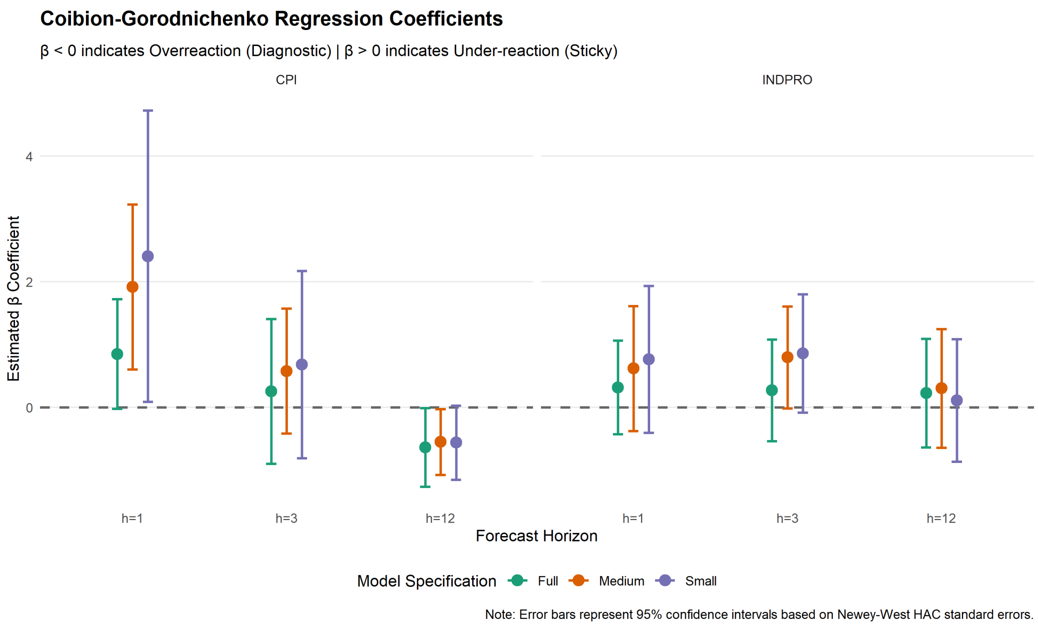

Figure 2: CG regression coefficients with confidence intervals

Notes: The figure plots $\hat{\beta}_h$ and normal-approximation confidence intervals based on Newey–West standard errors from Table 2.

To summarize the incremental effect of sentiment on the revision diagnostic, define $\Delta\beta_h = \beta_h^{Full} - \beta_h^{Small}$. Across horizons, $\Delta\beta_h$ is negative for CPI inflation, consistent with the full model moving the short-horizon coefficient toward zero, but the differences are not statistically distinguishable from zero in this sample (for CPI at $h = 1$, $\Delta\beta = -1.56$ with $p = 0.219$). Figure 3 visualizes these differences and corresponding uncertainty bands.

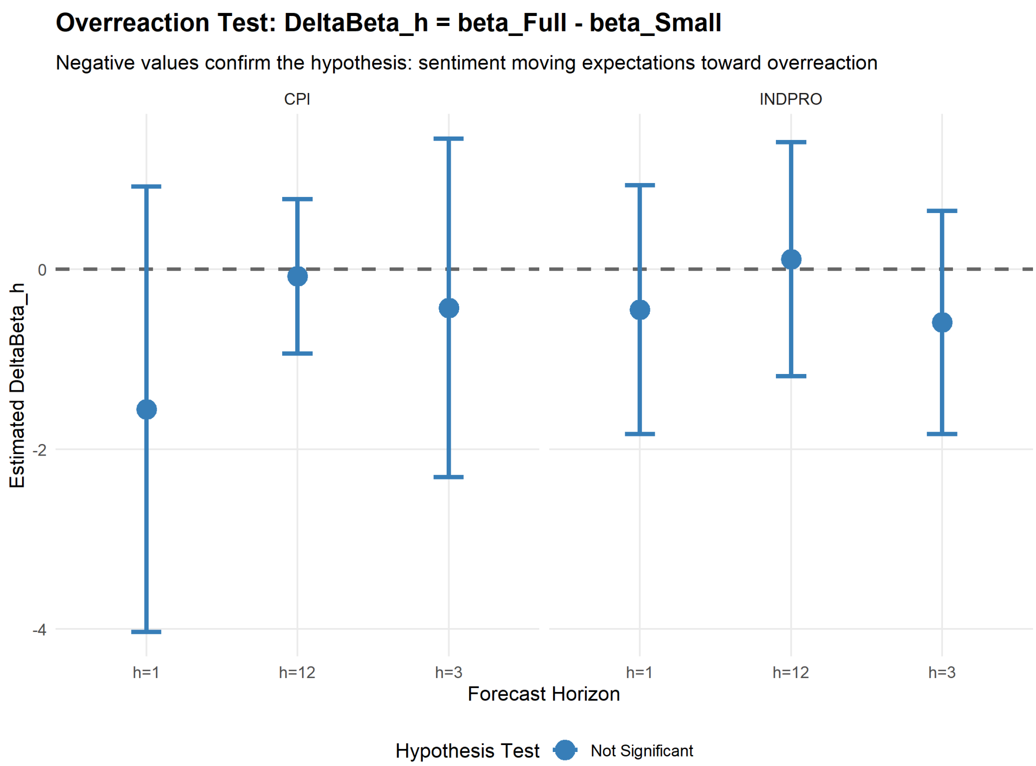

Figure 3: Difference in revision coefficients: Full minus Small

Notes: The figure is generated from `results/tables/delta_beta_overreaction_test.csv` and reports $\Delta\beta_h = \beta_h^{Full} - \beta_h^{Small}$ along with standard-error-based uncertainty bands.

4.4 Hyperparameter adaptation, shrinkage, and model size

The hierarchical setup makes it possible to quantify how the data select shrinkage as the information set expands. In the recursive estimation output, the posterior mean of the Minnesota tightness parameter λ is systematically lower for larger models: averaged across forecast origins, λ is 0.52 in the small model, 0.42 in the medium model, and 0.19 in the full model. This pattern is consistent with the role of hierarchical shrinkage in stabilizing estimation as dimensionality increases. Figure 4 plots the evolution of the posterior mean of λ over forecast origins.

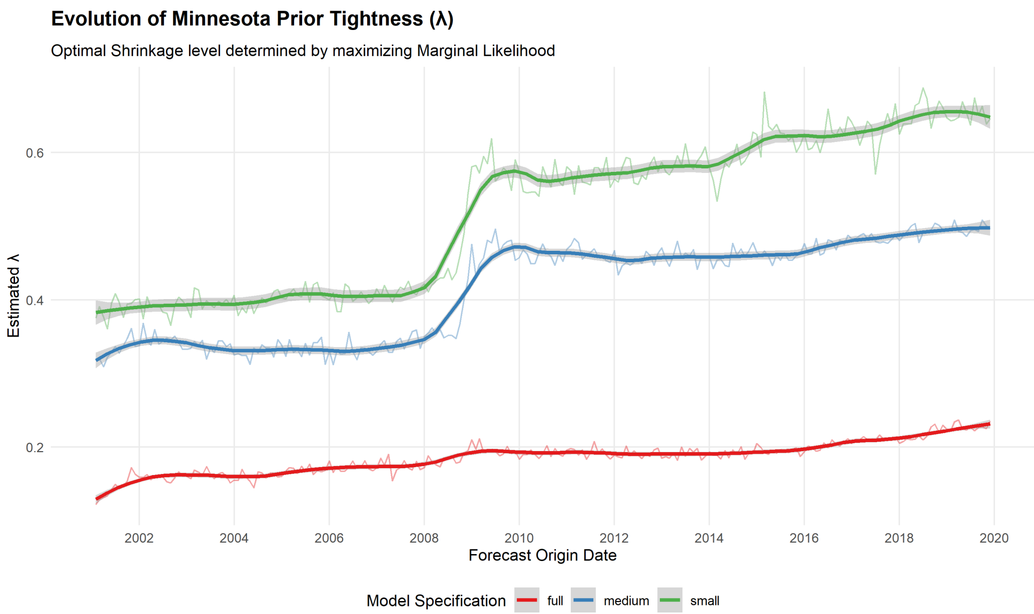

Figure 4: Evolution of hierarchical tightness parameter λ

Notes: The figure uses posterior means of λ recorded at each recursive forecast origin in `results/forecasts/hyperparameters_evolution.csv`.

4.5 Robustness and diagnostic decompositions

Two robustness exercises vary (i) the lag length from $p = 12$ to $p = 6$ and (ii) the initial training window endpoint from 2000M12 to 1995M12, keeping the rest of the design unchanged. The qualitative implications are stable: the medium specification remains the strongest performer for industrial production at short horizons, and the full specification remains competitive for inflation, including delivering the lowest twelve-month inflation RMSFE in the alternative initial-window design (1.286). Appendix Table 3 summarizes these results.

The output directory also includes diagnostics that describe the sources of forecast error. The MSE decomposition indicates that, for inflation, the bias component is small relative to variance and covariance components across horizons, while rolling relative RMSFEs highlight meaningful time variation in relative performance once a rolling window becomes available. These diagnostics help interpret why some gains appear concentrated in particular subperiods rather than uniformly across the evaluation sample.

4.6 Conclusion

The empirical evidence supports a nuanced version of Hypothesis 1. In this sample and evaluation design, the hierarchical BVAR improves inflation forecasts relative to a no-change benchmark across horizons, and sentiment contributes primarily at longer horizons. For real activity, financial prices deliver the most consistent incremental gains at short horizons, while longer-horizon industrial production forecasts remain challenging relative to a no-change benchmark. Hypothesis 2 also receives qualified support: for inflation, adding sentiment reduces the magnitude of the short-horizon revision coefficient, shifting it toward zero, while long-horizon coefficients are negative across specifications, consistent with overreaction in revisions. The corresponding differences across information sets are estimated imprecisely, suggesting that sharper identification would likely require either larger samples, alternative targets, or richer measures of expectations and sentiment.

References

References

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.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

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

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

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

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

Appendix

Table 3: Robustness: RMSFEs under alternative lag length and training window

Scenario	Model	Target	$h = 1$	$h = 3$	$h = 12$
$p = 6$	Small	CPI	3.451	2.639	1.291
$p = 6$	Medium	CPI	3.445	2.641	1.294
$p = 6$	Full	CPI	3.529	2.673	1.342
$p = 6$	Small	INDPRO	7.582	5.410	4.825
$p = 6$	Medium	INDPRO	7.336	5.033	4.621
$p = 6$	Full	INDPRO	7.494	5.173	4.572
Initial window ends 1995M12	Small	CPI	3.231	2.448	1.323
Initial window ends 1995M12	Medium	CPI	3.216	2.446	1.325
Initial window ends 1995M12	Full	CPI	3.267	2.431	1.286
Initial window ends 1995M12	Small	INDPRO	7.329	5.204	4.802
Initial window ends 1995M12	Medium	INDPRO	7.077	4.803	4.590
Initial window ends 1995M12	Full	INDPRO	7.218	4.930	4.453

Notes: Values are taken from `results/robustness/lag6/tables/rmsfe_results.csv` and `results/robustness/window1995/tables/rmsfe_results.csv`.