Forecast Alignment in High-Dimensional Linear Prediction: Factor-Based and Shrinkage Methods in Large Panels

Meishan Deng Maria Costanza Benedetti Filippo Belle

University of Bologna

Abstract

This papers studies whether commonly used high-dimensional forecasting methods extract similar predictive signals in large macroeconomic panels. We compare compression-based approaches (PCR and PLS) with shrinkage-based methods (Ridge, Lasso, and Elastic Net), focusing on the alignment of forecast paths. Monte Carlo simulations under increasingly general factor structures show that, despite distinct finite-sample behavior, forecast paths become highly correlated as the cross-sectional and time dimensions grow, reflecting the attenuation of weak spectral directions. An empirical application to U.S. macroeconomic data yields consistent evidence: forecasts are strongly aligned across methods even when predictive gains over a random walk are limited. These results highlight the central role of covariance structure in high-dimensional forecasting.

1 Introduction

Many problems in economics require the exploitation of large panels of time series. While the increasing availability of rich data sets offers the potential for more accurate and informative forecasts, it also gives rise to the well-known curse of dimensionality. Conventional multivariate regression methods often suffer from parameter non-identification, substantial variance inflation, and unstable out-of-sample performance due to overfitting. The linear forecasting solutions can broadly be classified into two categories. The first consists of factor-based or compression methods, such as principal component regression (PCR) and partial least squares (PLS), which forecast using a small number of low-dimensional latent components that capture the main features of the covariance structure of the predictors. The second comprises regularization or shrinkage methods, including Ridge, Lasso, and Elastic Net, which stabilize estimation by introducing penalty terms that trade off bias for reduced variance.

These methodologies often deliver very similar out-of-sample forecasts in macroeconomic applications, a pattern that is consistent with the strong collinearity observed in macroeconomic data. The related literature has established the asymptotic equivalence between Ridge and PCR. The seminal work of Bai (2003) demonstrates the consistency of PCR when the data exhibit an approximate factor structure and both the cross-sectional dimension n and the time dimension T grow large. Building on a weak factor structure, De Mol et al. (2008) show that when the cumulative impact of standardized factors on the cross section dominates the idiosyncratic components asymptotically, ridge regression achieves (n, T), consistent forecasting performance and empirically exhibits behavior close to factor-based methods in large panels. More recently, De Mol et al. (2024) consider a more general covariance spectral structure in which eigenvalues diverge at heterogeneous rates, establish the consistency of Ridge for the predictive component driven by dominant eigenvalue clusters, and show that ridge regression provides an effective alternative to principal component regression.

Building on these related literatures that establishes asymptotic equivalence between compression and shrinkage forecasts, this paper examines whether such equivalence extends beyond the canonical comparison between PCR and ridge. The underlying is that high-dimensional forecasting can be interpreted as approximating a low-dimensional predictive signal that resides in the dominant eigenspace of the

predictor covariance matrix. From this perspective, different forecasting methods may yield highly correlated predictions because they effectively recover the same dominant subspace, while their differences mainly reflect how weaker directions in the data are treated. We investigate this hypothesis by studying whether PCR, PLS, Ridge, Lasso, and Elastic Net generate highly correlated forecast paths in large panels of predictors, and by characterizing the data structures under which such correlations strengthen as both the cross-sectional dimension n and the time dimension T increase.

To this end, we conduct Monte Carlo simulations under a sequence of increasingly general factor-type environments, including the approximate factor model of Bai (2003), the weak-factor structure of De Mol et al. (2008), and the generalized eigenvalue-growth framework of De Mol et al. (2024). Across these designs, we systematically compare forecast correlations and relative predictive performance across methods, emphasizing the role of spectral separation between dominant and weak components. Then we complement the simulation evidence with an empirical study using U.S. macroeconomic data from 1961 to 2019, focusing on rolling forecasts of inflation and industrial production. The empirical results show that different high-dimensional methods produce highly correlated forecast paths, supporting the view that compression and shrinkage approaches are largely substitutable in large panel macroeconomics forecasting applications.

2 Methodology

General setup Let $X_t = (X_{1t}, \dots, X_{nt})'$ denote an n-dimensional predictor vector and y_{t+h} the target variable. We consider linear forecasting models of the form $y_{t+h} = X_t'\beta + u_{t+h}$, where n may be large relative to the available sample size T. In such environments, unrestricted least squares estimation is infeasible or unstable, motivating the use of dimension reduction or regularization techniques.

Principal component regression (PCR) PCR is an unsupervised compression method, as the projection directions are determined solely by the covariance structure of the predictors. By projecting the high-dimensional predictor vector onto the subspace spanned by the leading eigenvectors which explain the largest share of cross-sectional variation, PCR effectively applies a hard spectral truncation, assigning full weight to dominant components and zero weight to weaker directions.

Let $\Sigma_X = \mathbb{E}(X_t X_t')$ denote the covariance matrix of the predictors, with eigendecomposition

$$\Sigma_X = V \Lambda V',$$

where $\Lambda = \mathrm{diag}(\mu_1, \dots, \mu_n)$ with $\mu_1 \geq \cdots \geq \mu_n$, and $V = (v_1, \dots, v_n)$ collects the corresponding eigenvectors. Let $V_r = (v_1, \dots, v_r)$ denote the matrix of the leading r eigenvectors. Principal component regression projects the predictors onto the associated low-dimensional subspace,

$$Z_t^{\mathrm{PCR}} = V_r' X_t,$$

which coincides with the first r principal components of X_t. Forecasts are then obtained by regressing y_{t+h} on the projected predictors,

$$\hat{y}_{t+h}^{\mathrm{PCR}} = Z_t^{\mathrm{PCR},\prime} \hat{\beta} = X_t' V_r \hat{\beta},$$

where $\hat{\theta}$ is estimated by ordinary least squares.

Partial least squares (PLS) PLS is a supervised compression method, as the projection directions are selected by using information from the forecasting target. It projects the high-dimensional predictor vector onto a low-dimensional subspace that maximizes predictive relevance, placing relatively larger weight on directions that are informative for forecasting. This focus on predictive covariance can reduce bias in the presence of weak but relevant signals, at the cost of potentially higher estimation variance.

Let $W_r = (w_1, \dots, w_r)$ denote the matrix collecting the first r PLS weight vectors, where each w_j is chosen to maximize the covariance between $w_j' X_t$ and y_{t+h}, subject to orthogonality conditions across

components. The predictors are projected onto the associated low-dimensional subspace,

$$Z_t^{\text{PLS}} = W_r' X_t,$$

yielding the first r PLS components. Forecasts are then obtained by regressing y_{t+h} on the projected predictors,

$$\hat{y}_{t+h}^{\text{PLS}} = Z_t^{\text{PLS},\prime} \hat{\beta} = X_t' W_r \hat{\beta},$$

where $\hat{\beta}$ is estimated by ordinary least squares.

Relative to PCR, the projection directions in PLS need not coincide with the eigenvectors of the predictor covariance matrix. However, in environments where the same latent factor both drives the dominant eigenvalue of $\text{Cov}(X_t)$ and constitutes the source of predictability for y_{t+h}, the PLS and PCR subspaces tend to align, leading to highly correlated forecast paths.

Ridge regression Ridge regression is a supervised shrinkage-based forecasting method that introduces an ℓ_2 penalty to the least squares objective, shrinking regression coefficients continuously toward zero. By trading off bias for reduced estimation variance, ridge regression stabilizes coefficient estimates.

Formally, the ridge estimator is defined as

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta} \sum_t (y_{t+h} - X_t'\beta)^2 + \lambda \|\beta\|_2^2,$$

where $\lambda > 0$ is a regularization parameter.

Relative to PCR, ridge regression retains all directions in the predictor space but applies continuous shrinkage through a penalty parameter. In high-dimensional settings with strong collinearity, this yields an implicit spectral weighting in which directions associated with smaller eigenvalues receive progressively less weight. When predictive content is concentrated in a dominant eigenspace and regularization strength increases with dimensionality, ridge effectively emphasizes the same leading directions selected by PCR, resulting in highly correlated forecast paths.

Lasso regression Lasso is a supervised shrinkage-based forecasting method that introduces an ℓ_1 penalty, inducing sparsity by shrinking some coefficients exactly to zero and thereby concentrating predictive weight on a limited set of predictors.

Formally, the lasso estimator is defined as

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \sum_t (y_{t+h} - X_t'\beta)^2 + \lambda \|\beta\|_1,$$

where $\lambda > 0$ is a regularization parameter.

Lasso is most effective when the regression coefficient is approximately sparse in the original predictor coordinates. In factor-driven panels with strong collinearity, however, the predictive component is often dense at the variable level even when it is low-dimensional in a factor sense, so lasso may select a small set of proxy variables whose identity varies across samples.

Relative to PCR and ridge, forecast alignment of lasso therefore depends on whether the fitted signal lies in the same predictive subspace and on how the penalty level is chosen. As the predictive subspace becomes more precisely identified, lasso forecasts can become increasingly aligned with those from spectral and shrinkage methods, even if the selected variables remain unstable.

Elastic Net regression Elastic Net is a supervised shrinkage-based forecasting method that combines ℓ_1 and ℓ_2 penalties, allowing for both coefficient shrinkage and variable selection. By blending smooth shrinkage with sparse selection, elastic net not only performs variable selection but also accommodates groups of strongly correlated predictors.

Formally, the elastic net estimator is defined as

$$\hat{\beta}^{\mathrm{EN}} = \arg\min_{\beta} \sum_{t} (y_{t+h} - X_t'\beta)^2 + \lambda \left[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right],$$

where $\lambda > 0$ controls the overall strength of regularization and $\alpha \in [0,1]$ governs the relative importance of sparsity versus shrinkage. Setting $\alpha = 0$ yields ridge regression, while $\alpha = 1$ corresponds to lasso.

Elastic net is particularly useful when predictive information is distributed across groups of correlated predictors while additional variables are largely uninformative. By combining ℓ_1 and ℓ_2 penalties, it allows irrelevant predictors to be excluded while stabilizing estimation across correlated variables, which can yield forecast paths that differ from both ridge and lasso depending on how the data balance sparsity and collinearity. Its correlation with PCR forecasts depends on whether elastic net places effective weight on variance-dominant directions in the predictor space.

3 Monte Carlo Simulation

3.1 Simulation Design

We consider a high-dimensional predictive regression in which all forecasting power is driven by a single latent factor. Let $X_t \in \mathbb{R}^n$ denote the predictor panel at time t, and let y_{t+h} be the target variable observed h periods ahead. Across all designs, the predictive relation is fixed as

$$y_{t+h} = \gamma F_t + v_{t+h}, \qquad \gamma \neq 0,$$

so that the one-dimensional latent factor F_t is the unique source of predictability. By construction, no other component of X_t enters the conditional mean of y_{t+h}. Consequently, differences in forecasting performance across designs are entirely attributable to the difficulty of estimating the predictive direction from a noisy high-dimensional predictor panel, rather than to changes in the signal itself.

The predictor panel admits the factor representation

$$X_t = \Lambda F_t + \sum_{j=1}^{r_n} \Lambda_j G_{j,t} + \xi_t,$$

where F_t is the predictive factor, $G_{j,t}$ are non-predictive common noise factors, and ξ_t is an idiosyncratic component. The predictive factor F_t, the noise factors $G_{j,t}$, and the forecast innovations v_t are independently generated as i.i.d. standard Gaussian processes.

The loading vector Λ is drawn with i.i.d. Gaussian entries and normalized such that $\frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 = 1$, which implies $\|\Lambda\|^2 \asymp n$ and ensures that the signal component corresponds to a strong factor generating a dominant eigenvalue of order $O(n)$ in the covariance of X_t.

The idiosyncratic component ξ_t is Gaussian with covariance matrix $\Psi = DR(\rho_{\mathrm{cs}})D$, where $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \overset{iid}{\sim} U[0.6, 1.4]$, and $R(\rho_{\mathrm{cs}})$ is a Toeplitz correlation matrix with entries $\rho_{\mathrm{cs}}^{|i-j|}$ and $\rho_{\mathrm{cs}} = 0.2$. This specification introduces weak but non-negligible cross-sectional dependence while ruling out additional strong factors. All components are serially independent over time.

We fix the forecast horizon to $h = 1$. The cross-sectional and time dimension ranges over $n \in \{5, 10, 20, 30, 40, 50, 80, 100, 200, 300, 500, 1000\}, T \in \{80, 80, 80, 80, 80, 80, 80, 100, 200, 300, 500, 1000\}$. Estimation is conducted using a time-ordered split, with the last 20% of observations reserved for out-of-sample evaluation and a minimum training window of 40 observations. For each design and each (n, T) pair, results are averaged over $B = 200$ independent Monte Carlo replications.

3.2 Evaluation criteria

We evaluate forecasting performance and asymptotic behavior using three complementary criteria: estimation loss relative to the oracle based on out-of-sample mean squared error, convergence rates, and forecast alignment with principal component regression.

As a lower bound on achievable forecast error under the data-generating process, we construct an oracle benchmark that exploits knowledge of the true predictive factor F_t. On the training sample $t = 1, \dots, T_{\text{tr}}$, the oracle regression coefficient is estimated as

$$\widehat{\gamma}_{\text{or}} = \frac{\sum_{t=1}^{T_{\text{tr}}} F_t \, y_{t+1}}{\sum_{t=1}^{T_{\text{tr}}} F_t^2},$$

and oracle forecasts are given by $\widehat{y}_{t+1}^{\text{or}} = \widehat{\gamma}_{\text{or}} F_t$ for $t = T_{\text{tr}} + 1, \dots, T^*$, where $T^* = T - 1$. The corresponding oracle test mean squared error is

$$\text{MSE}_{\text{or}} = \frac{1}{T_{\text{te}}} \sum_{t=T_{\text{tr}}+1}^{T^*} \left(y_{t+1} - \widehat{y}_{t+1}^{\text{or}}\right)^2, \qquad T_{\text{te}} = T^* - T_{\text{tr}}.$$

Because the oracle forecast uses the true predictive factor, MSE_{or} represents the benchmark error that would obtain if the predictive direction were known and only the scalar coefficient had to be estimated.

For any forecasting method m producing test forecasts $\widehat{y}_{t+1}^{(m)}$, we compute the out-of-sample mean squared error

$$\text{MSE}_m = \frac{1}{T_{\text{te}}} \sum_{t=T_{\text{tr}}+1}^{T^*} \left(y_{t+1} - \widehat{y}_{t+1}^{(m)}\right)^2.$$

To isolate the loss arising from estimating the predictive direction in high dimensions, we define the estimation loss

$$\mathcal{L}_m = \Delta MSE = \text{MSE}_m - \text{MSE}_{\text{or}}.$$

By construction, $\mathcal{L}_m \geq 0$, and $\mathcal{L}_m \to 0$ if and only if method m asymptotically recovers the true predictive factor. This quantity therefore directly measures the extent to which a method succeeds in identifying the correct predictive direction, abstracting from irreducible forecast noise.

Then, we study asymptotic behavior by estimating convergence rates as dimensionality increases. Let $Q_m(n, T)$ denote a generic nonnegative error measure. For each grid point (n, T), the error measure is first averaged across Monte Carlo replications. Convergence rates are then summarized using log–log regressions based on these aggregated quantities. Specifically, we estimate marginal convergence rates using separate regressions of the form

$$\log Q_m(n) = \alpha_m + \beta_m^{(n)} \log n + \varepsilon_m,$$

and

$$\log Q_m(T) = \tilde{\alpha}_m + \beta_m^{(T)} \log T + \tilde{\varepsilon}_m,$$

where in the second regression T is taken to be the average training sample size associated with each cross-sectional dimension, denoted by T_{mean}.

The coefficients $\beta_m^{(n)}$ and $\beta_m^{(T)}$ therefore capture the marginal rates at which the error measure decays as the cross-sectional dimension or the time dimension increases, respectively. In our main simulation grid, T grows roughly proportionally with n, so the two marginal slopes are typically similar. Negative values of the estimated slopes indicate that the estimation loss or directional error vanishes as dimensionality increases, while larger magnitudes of $|\beta_m|$ correspond to faster recovery of the predictive direction.

Finally, to assess directional agreement across data-driven procedures, we further report forecast alignment with principal component regression. Let $\widehat{y}_{t+1}^{\mathrm{pcr}}$ denote the PCR forecast constructed using the number of factors selected by the Bai–Ng information criterion. For each method m, forecast alignment is measured by

$$\rho_{m,\mathrm{pcr}} = \mathrm{corr}\big(\widehat{y}_{t+1}^{(m)}, \widehat{y}_{t+1}^{\mathrm{pcr}}\big),$$

computed on the test sample. High values of $\rho_{m,\mathrm{pcr}}$ indicate that different methods extract similar predictive components from the predictor panel, even when their finite-sample forecast accuracy differs.

3.3 Factor Structure Designs

The factor structure designs differ only in the cross-sectional structure of the non-predictive components in X_t, while keeping the predictive relation fixed. In all cases, the predictive factor F_t is the unique source of predictability for y_{t+h}, and differences across designs operate exclusively through the spectral properties of the predictor covariance matrix $\Sigma_X = \mathbb{E}(X_t X_t')$.

Approximate factor structure (AFS) In the AFS benchmark, the predictor panel contains a single strong factor, coinciding with the predictive factor. The loading vector Λ is normalized so that

$$\|\Lambda\|^2 \asymp n,$$

which implies a dominant eigenvalue

$$\mu_1(\Sigma_X) = O(n),$$

while all remaining eigenvalues are bounded. This environment features a clear spectral separation, under which principal component–based methods consistently recover the predictive direction.

Weak factor structure (WFS) The WFS design extends AFS by introducing two non-predictive common noise factors with sparse cross-sectional impact. Each noise factor loads only on a subset of size $m_n = n^\delta$, with $0 < \delta < 1$. The covariance spectrum satisfies

$$\mu_1(\Sigma_X) = O(n), \qquad \mu_j(\Sigma_X) = O(n^\delta), \quad j \geq 2.$$

The predictive factor remains uniquely strong, but the leading eigenspace is contaminated by moderately diverging non-predictive directions.

Generalized factor structure (GFS) The GFS further relaxes WFS by allowing heterogeneous degrees of cross-sectional pervasiveness among non-predictive factors. Each noise factor $G_{j,t}$ is associated with a loading vector Λ_j satisfying

$$\|\Lambda_j\|^2 \asymp n^{\alpha_j}, \qquad 0 < \alpha_j < 1,$$

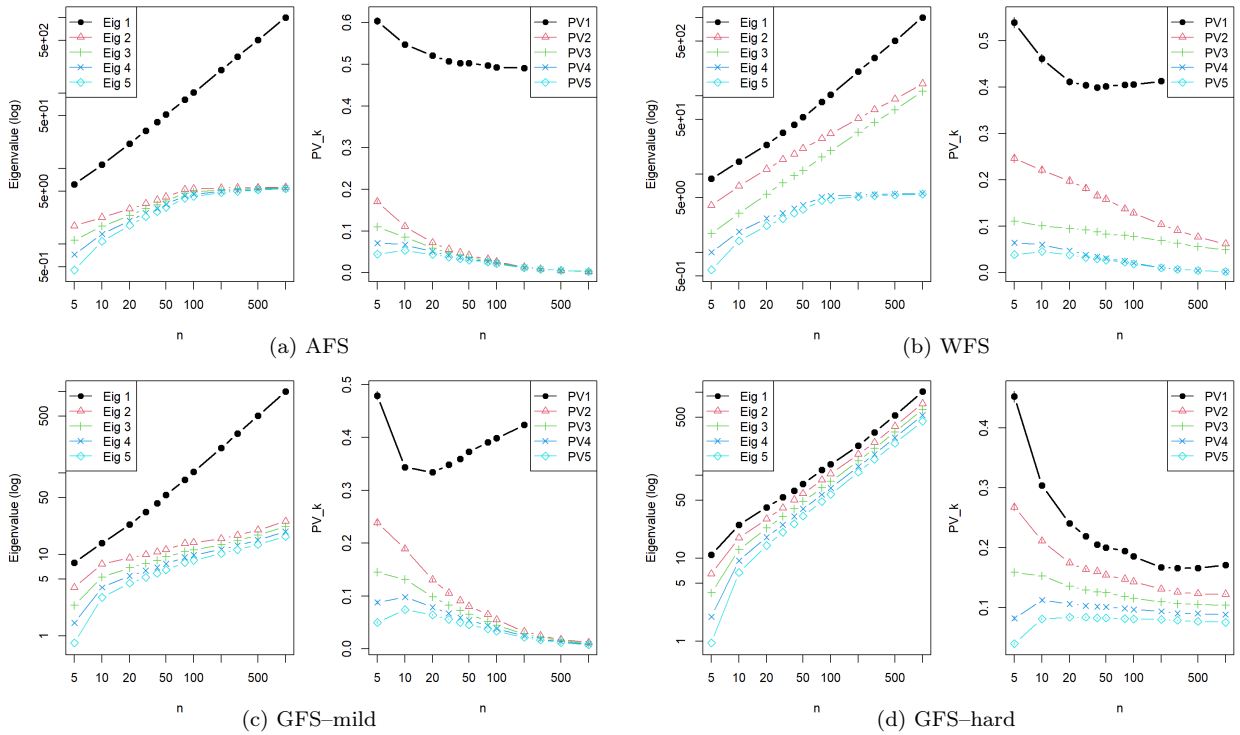
implying sublinearly diverging eigenvalues

$$\mu_{j+1}(\Sigma_X) \asymp n^{\alpha_j}.$$

To simplify, the number of noise factors r_n is fixed at ten, but the dominant eigenspace thickens as n grows. Two regimes are considered. In GFS-mild, $\alpha_j \in [0.25, 0.45]$, generating moderate spectral contamination. In GFS-hard, $\alpha_j \in [0.75, 0.95]$, so noise eigenvalues grow close to the strong-factor rate, yielding a dense dominant subspace in which the predictive direction is weakly identifiable.

Figure 1 illustrates the spectral properties of the predictor covariance matrix across the four factor structure designs. In the AFS case, the spectrum is dominated by first eigenvalue that grows proportionally with the cross-sectional dimension, while all remaining eigenvalues remain bounded, indicating a clear separation between the predictive factor and the idiosyncratic components. Under WFS, two additional eigenvalues diverge at a sublinear rate relative to the leading eigenvalue, thereby weakening

Figure 1: Spectral properties of the predictor covariance across factor structure designs



(a) AFS (b) WFS

(c) GFS–mild (d) GFS–hard

Panels (a)–(d) report the leading eigenvalues of the sample covariance matrix of the predictors (left) and the corresponding proportions of variance explained by the first five principal components (right) for the AFS, WFS, GFS–mild, and GFS–hard designs, respectively. Eigenvalues and variance shares are computed from the training sample.

spectral separation. In the GFS–mild design, several non-predictive factors generate multiple sublinearly diverging eigenvalues with heterogeneous growth rates, which remain well separated from the leading one so that the dominant eigenspace is still relatively distinct. By contrast, in the GFS–hard design, several eigenvalues diverge at heterogeneous but near-linear rates, resulting in a much smaller separation among the leading eigenvalues and a comparatively dense dominant spectrum.

3.4 Simulation Results

Figure 2 reports forecast correlations between alternative methods and PCR. A common pattern emerges across all designs: forecast correlations increase with the cross-sectional dimension (n) and approach unity in moderate to large samples. This convergence is not confined to the canonical comparison between PCR and Ridge, but extends to PLS, Lasso, and Elastic Net.

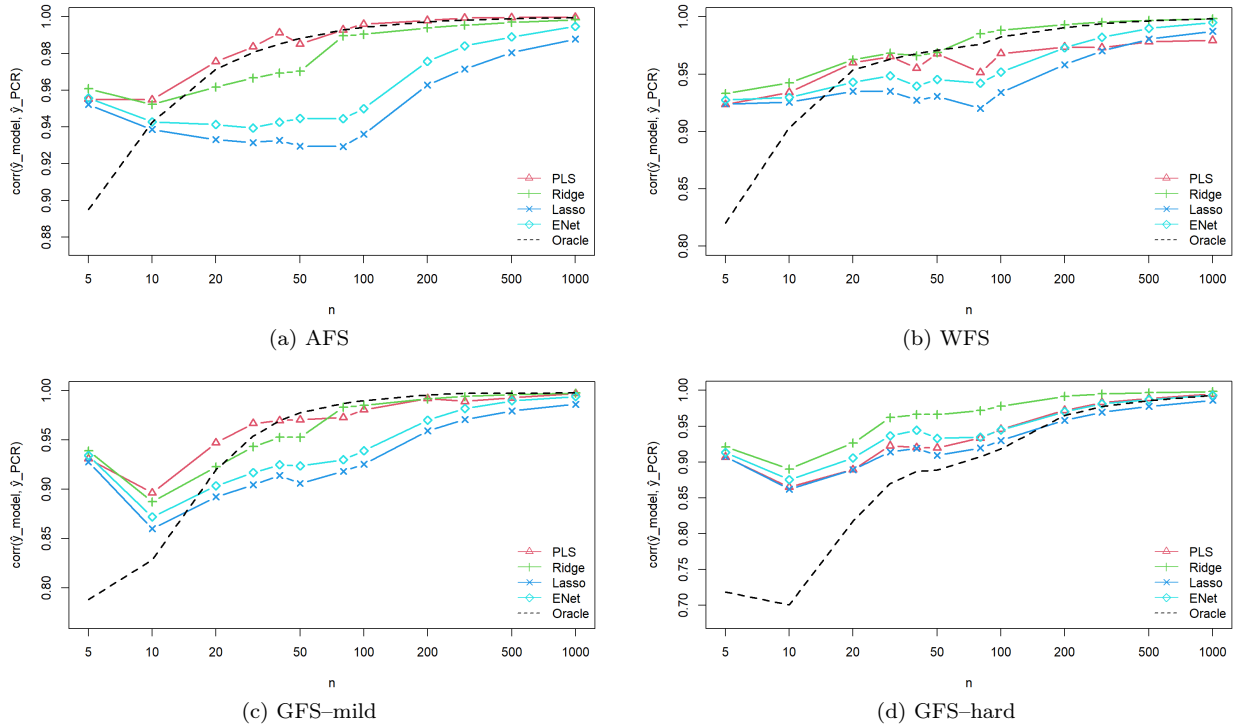
Forecast alignment across methods depends critically on how different procedures adjust their effective dimensionality as the cross-sectional dimension increases. In environments with clear spectral separation, such as the AFS and GFS–mild designs, the predictive signal is concentrated in a small number of dominant eigenvectors. PCR and PLS therefore select similarly low-dimensional representations of the predictor space, leading to the tightest forecast alignment among all methods. In these settings, both procedures effectively recover the same leading predictive direction, with PLS benefiting from its supervised weighting when the signal-to-noise contrast is strong.

As spectral contamination increases, the mechanism generating alignment shifts. Under the WFS and GFS–hard designs, weak or near-linearly diverging noise factors blur the distinction between predictive and non-predictive components. PCR responds by selecting an increasing number of components as n grows—expanding from a small set of dominant factors under WFS to a much broader subspace under GFS–hard that includes both the predictive factor and common noise directions. At the same time,

Ridge regression increases its effective shrinkage strength with dimensionality, progressively averaging information across the same dominant eigenspace. These endogenous adjustments cause PCR and Ridge to converge toward similar linear combinations of predictors, generating highly correlated forecast paths. This pattern is consistent with the asymptotic equivalence results of De Mol et al. (2008) and De Mol et al. (2024).

Lasso and Elastic Net exhibit weaker alignment in smaller samples, reflecting the instability induced by sparse selection when predictive information is diffusely distributed. As dimensionality increases, however, cross-validated penalties initially enforce strong sparsity but subsequently relax, leading to a growing number of selected predictors and increasingly dense fitted models. In the GFS–hard design, where predictive content is spread across many weakly separated directions, this transition causes Lasso and Elastic Net to approximate ridge-type shrinkage. Consequently, differences in forecast alignment across all five methods largely disappear, with correlations converging to similarly high levels. Because such dense and weakly separated spectral structures resemble those observed in large macroeconomic panels, these results suggest that, in empirically relevant settings and sufficiently large samples, compression and shrinkage methods are likely to produce strongly correlated forecasts.

Figure 2: Forecast alignment with principal component regression



(a) AFS (b) WFS

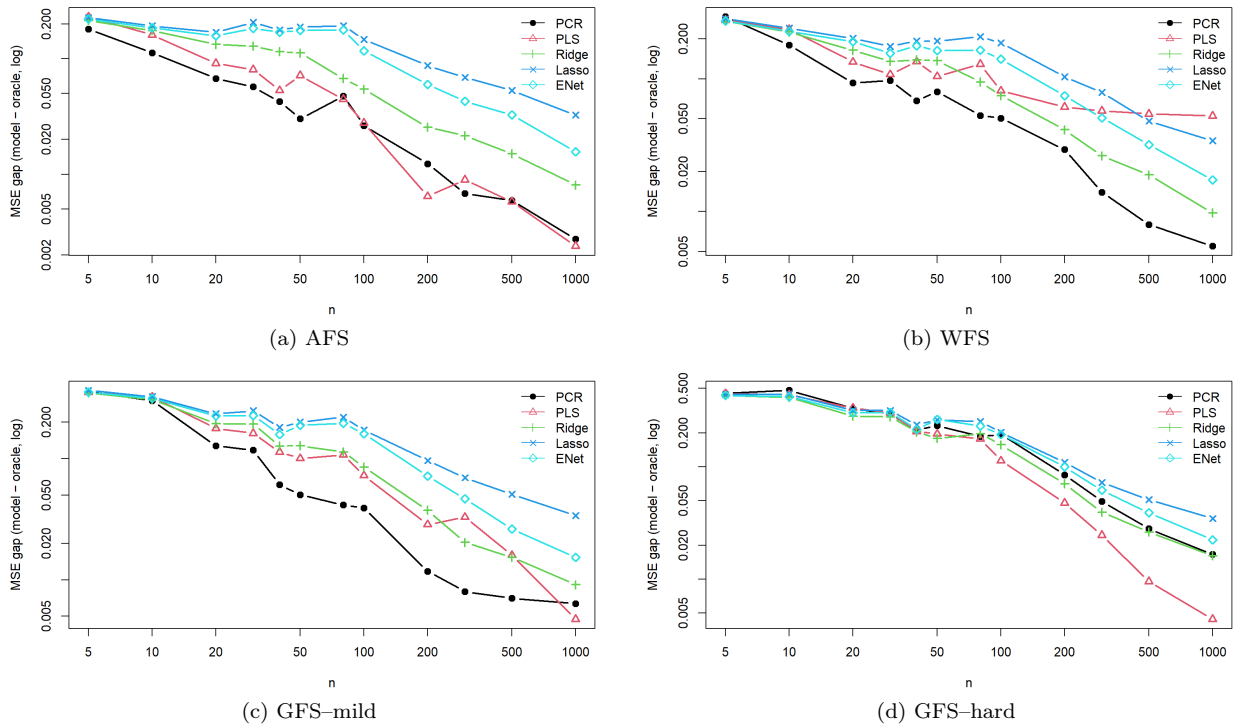
(c) GFS–mild (d) GFS–hard

Panels (a)–(d) report forecast alignment with principal component regression, measured by the out-of-sample correlation between the forecasts produced by each method and those from PCR, across the AFS, WFS, GFS–mild, and GFS–hard designs, respectively. Correlations are computed on the test sample and averaged across Monte Carlo replications.

Figure 3 reports estimation loss relative to the oracle. Estimation loss declines monotonically with the cross-sectional dimension n in all designs, but this monotonic decline does not imply that all methods eliminate noise directions as dimensionality increases. In particular, PCR retains common noise components. Besides, the level of the loss and the ranking of methods vary substantially across spectral environments. These design-dependent level differences motivate a complementary analysis of convergence speeds to provide interpretation for the cross-structure forecast correlation patterns above.

Table 1 shows the convergence speeds as n and T increase. Convergence arises through different mechanisms across methods. For PCR, convergence behavior is governed by how variance in the predictor

Figure 3: Estimation loss relative to the oracle across factor structure designs



(a) AFS (b) WFS

(c) GFS–mild (d) GFS–hard

Panels (a)–(d) report the estimation loss relative to the oracle, defined as the difference between the out-of-sample mean squared error of each method and that of the oracle forecast, across the AFS, WFS, GFS–mild, and GFS–hard designs, respectively. Losses are averaged across Monte Carlo replications and plotted against the cross-sectional dimension n.

space is distributed across predictive and non-predictive directions. When the leading eigenvalue is clearly separated, as in AFS, or when additional noise eigenvalues diverge only sublinearly and remain well separated, as in WFS and GFS–mild, most of the variance in X_t is concentrated along a small number of directions. In these environments, the Bai–Ng criterion selects a low and stable number of principal components, so the predictive factor is isolated early and the forecasting regression involves only a limited set of noise-related components. Estimation error then declines at a similar pace across these designs as sample size increases. However, In GFS–hard, several non-predictive factors generate eigenvalues that grow at rates close to that of the predictive factor, so variance no longer clearly distinguishes signal from noise. As a result, the Bai–Ng criterion retains a much larger set of components, forcing PCR to estimate coefficients on many common-noise directions. Although these coefficients are asymptotically zero, their gradual attenuation requires increasing sample information, resulting in the weak convergence rates despite its asymptotic consistency.

Table 1: Convergence rates of the estimation loss relative to the oracle

	Factor-structure design			
Methodology	AFS	WFS	GFS–mild	GFS–hard
PCR	-0.78	-0.75	-0.88	-0.66
PLS	-0.89	-0.33	-0.78	-0.92
Ridge	-0.65	-0.64	-0.74	-0.66
Lasso	-0.36	-0.39	-0.45	-0.51
Elastic Net	-0.49	-0.50	-0.60	-0.57

In contrast, the supervised nature of PLS makes its convergence sensitive to the stability of predictive covariance, and this sensitivity is amplified by cross-validation. PLS uses predictive covariance to rank candidate directions, while cross-validation selects among these directions based on out-of-sample

performance. Under WFS, where a small number of weak noise factors have fixed cross-sectional impact, these factors can generate persistent but spurious predictive covariance in finite samples, thereby cross-validation fails to reliably discriminate them from the true predictive direction. PLS therefore fluctuates between the true predictive direction and noise-driven directions, leading to particularly slow convergence. Under GFS–hard, by contrast, noise factors are numerous and heterogeneous, and no individual noise direction can sustain stable predictive covariance across sample splits. Spurious directions therefore perform inconsistently out of sample and are penalized by cross-validation, while the uniquely persistent predictive factor consistently dominates. This allows PLS to rapidly concentrate on the true predictive direction, resulting in the fastest convergence rates among the methods considered.

Shrinkage-based methods, Ridge, Lasso, and Elastic Net, exhibit heterogeneous but comparatively stable convergence behavior. Although their tuning paths differ, all three procedures adjust hyper-parameter continuously rather than making discrete dimension or direction selections, resulting in their convergence behavior relatively insensitive to variations in spectral complexity. For Ridge, cross-validated penalties increase with dimensionality, strengthening continuous shrinkage and rapidly attenuating the influence of weak directions. The Lasso penalty typically decreases with n, leading to the selection of an expanding set of predictors, while Elastic Net shifts toward ridge-type regularization as α declines and λ increases.Despite these distinct adjustment paths, all three methods converge toward similar linear combinations of predictors.

Notably, shrinkage methods exhibit faster convergence under GFS–hard than under AFS. This difference reflects how informative the underlying spectral environment is for regularization. In AFS, the predictive factor is sharply separated from noise, and regularization plays a limited role beyond stabilizing estimation. Because alternative regularization paths deliver similar out-of-sample performance, cross-validated penalties adjust only modestly, and convergence relies primarily on sample averaging. By contrast, in GFS–hard, predictive information is dispersed across many weakly separated and heterogeneous directions. This heterogeneity prevents any single noise component from generating persistent pseudo-predictive structure and makes poorly regularized models perform systematically worse out of sample. As a result, differences across regularization paths become large and stable, providing strong identification signals for cross-validation.

4 Empirical Study

The data used in the out-of-sample forecasting analysis are taken from the monthly U.S. macroeconomic database from the federal reserve bank of St. Louis. The panel includes real variables, nominal variables, asset prices, and, for a total of $n = 126$ variables. Series are transformed at the annual frequency to obtain stationarity, followed by McCracken and Ng (2016). Real variables are expressed in annual growth rates, while variables already expressed in rates are transformed using annual first differences. Price variables are expressed in annual inflation rates.

The diagnostic inspection of the transformed series reveals the presence of extremely large observations in the series measuring nonborrowed reserves of depository institutions (NONBORRES), particularly during the period of unconventional monetary policy. Given the magnitude of these observations relative to the scale of the series, as shown in Figure 13, this variable is excluded from the baseline dataset. In addition, a small number of series with substantial missing observations are removed to preserve the largest possible time dimension. Finally, to avoid the structural break associated with the COVID-19 period, observations after December 2019 are excluded. The resulting dataset spans the period from 1961:01 to 2019:12 and contains 120 monthly macroeconomic variables.

Let $x_t \in \mathbb{R}^n$ denote the vector of standardized macroeconomic predictors available at month t, with $n = 118$. We consider two forecasting targets: industrial production and inflation, which are defined as $ip_t = 100 \log(IP_t)$, and $\pi_t = 100 \log(CPI_t/CPI_{t-12})$. For each target, we forecast $h = 12$ step

changes in the predictive regression, $w_{t+h}^{(i)} = y_{t+h} - y_t$, where $i \in \{ip, \pi\}$. The forecasts for the level of IP and CPI are recovered as $\widehat{y}_{t+h|t} = y_t + \widehat{w}_{t+h|t}^{(i)}$. As a benchmark, we use the random walk forecast, $\widehat{y}_{t+h|t}^{RW} = y_t$. For all methods we report results for $\rho = 0$ (no lags of the regressor), which is the one typically considered in macroeconomic applications.

We implement a monthly rolling forecasting scheme with a fixed estimation window of $W = 120$ months. At each forecast origin T_0, models are estimated using observations in the interval $[T_0 - W + 1, T_0]$, with predictors standardized using only in-sample information. Tuning parameters are selected via a nested time-series cross-validation procedure conducted within each window[1]. Specifically, we choose $K = 5$ validation anchors subject to a minimum spacing of 12 months. Because the target is w_{t+h}, the training sample for a validation anchor τ ends at $\tau - h$, ensuring that the validation target $w_{\tau+h}$ is strictly out-of-sample. For each candidate hyperparameter configuration, we compute the average validation mean squared error across anchors and select the window-specific optimum. The final forecast is then obtained by re-estimating the model on the full window using the selected hyperparameters.

For each method m, we compute the correlation between its level forecast path and that of PCR, $\rho_{m,pcr} = \mathrm{Corr}\Big(\widehat{y}_{t+h|t}^{m}\ ,\ \widehat{y}_{t+h|t}^{pcr}\Big)$, and report this measure for the full sample as well as for subsamples corresponding to distinct macroeconomic regimes (1971–1984, 1985–2007, and 2008–2019). To study how forecast similarity evolves over time, we additionally report both rolling correlations and expanding correlations computed recursively over the evaluation period.

Forecast accuracy is evaluated relative to a random walk benchmark. Specifically, we report the mean squared forecast error (MSFE) of each model, which is computed on level forecasts as

$$\mathrm{MSFE}(m) = \frac{1}{T_1 - T_0 - h + 1} \sum_{t=T_0}^{T_1 - h} \Big(y_{t+h} - \widehat{y}_{t+h|t}^{m}\Big)^2, \qquad y_t \in \{\pi_t, ip_t\}.$$

Finally, we summarize the distributions of selected hyperparameters across subsamples. These distributions document how model complexity varies across targets and macroeconomic regimes.

Table 2: Mean squared forecast error relative to the random walk

Period	PCR	PLS	Ridge	Lasso	Elastic Net
Panel A: Inflation (YoY)					
1971–2019	1.78	1.69	1.30	1.09	1.26
1971–1984	1.12	0.98	1.06	0.80	0.98
1985–2007	1.89	2.56	1.60	1.38	1.57
2008–2019	3.35	2.70	1.63	1.56	1.68
Panel B: Industrial Production (log level)					
1971–2019	1.44	1.12	1.06	1.28	1.29
1971–1984	1.29	0.66	0.70	0.71	0.84
1985–2007	1.17	1.17	0.92	0.99	0.97
2008–2019	2.17	2.12	2.07	3.00	2.77

Table 2 reports mean squared forecast errors (MSFE) relative to a random walk benchmark. For both targets, forecast performance exhibits pronounced time variation: in later subsamples, especially after 2008, relative MSFE increases substantially, and almost all methods fail to outperform the random walk. Forecast accuracy also tends to move together across methods: subsamples in which PCR performs relatively well are also those in which other methods achieve lower relative MSFE.

The relative performance of different methods also differs across the two forecasting targets. For inflation, Lasso delivers the lowest relative MSFE across subsamples. This suggests that predictive

[1] Cross-validation does not guarantee MSFE-minimizing forecasts ex post. Rather, it imposes a common, data-driven tuning rule that aligns the bias–variance trade-off across models.

information for inflation is more likely to be concentrated in a small number of indicators that are highly collinear with the target. In contrast, for industrial production, Ridge regression exhibits more effective performance. As a level measure of real economic activity, industrial production is highly collinear with a large set of macroeconomic indicators, and its predictive content is therefore more likely to be dispersed across many correlated variables.

Consistent with the simulation results, shrinkage-based methods (Ridge, Lasso, and Elastic Net) produce smoother forecasts in most subsamples than compression-based methods (PCR and PLS). PCR performs relatively weakly for both targets, as dimension reduction based on fixed component truncation is more sensitive to changes in the macroeconomic structure and may amplify forecast error volatility in finite samples, potentially due to the discrete nature of component selection under cross-validation. PLS exhibits the largest variation across targets, performing substantially better for industrial production than for inflation. A plausible interpretation is that, for industrial production, the covariance between the predictors and the target reflects more stable comovement in real activity, whereas for inflation such covariance is weaker and more sample-dependent, making supervised dimension reduction less reliable out of sample.

Table 3: Correlation of forecasts with PCR

Period	PLS	Ridge	Lasso	Elastic Net
Panel A: Inflation (YoY)				
1971–2019	0.77	0.90	0.87	0.89
1971–1984	0.87	0.91	0.85	0.89
1985–2007	0.53	0.67	0.51	0.61
2008–2019	0.20	0.69	0.66	0.68
Panel B: Industrial Production (log level)				
1971–2019	0.99	0.99	0.99	0.99
1971–1984	0.87	0.90	0.90	0.91
1985–2007	0.99	0.99	0.99	0.99
2008–2019	0.91	0.86	0.90	0.88

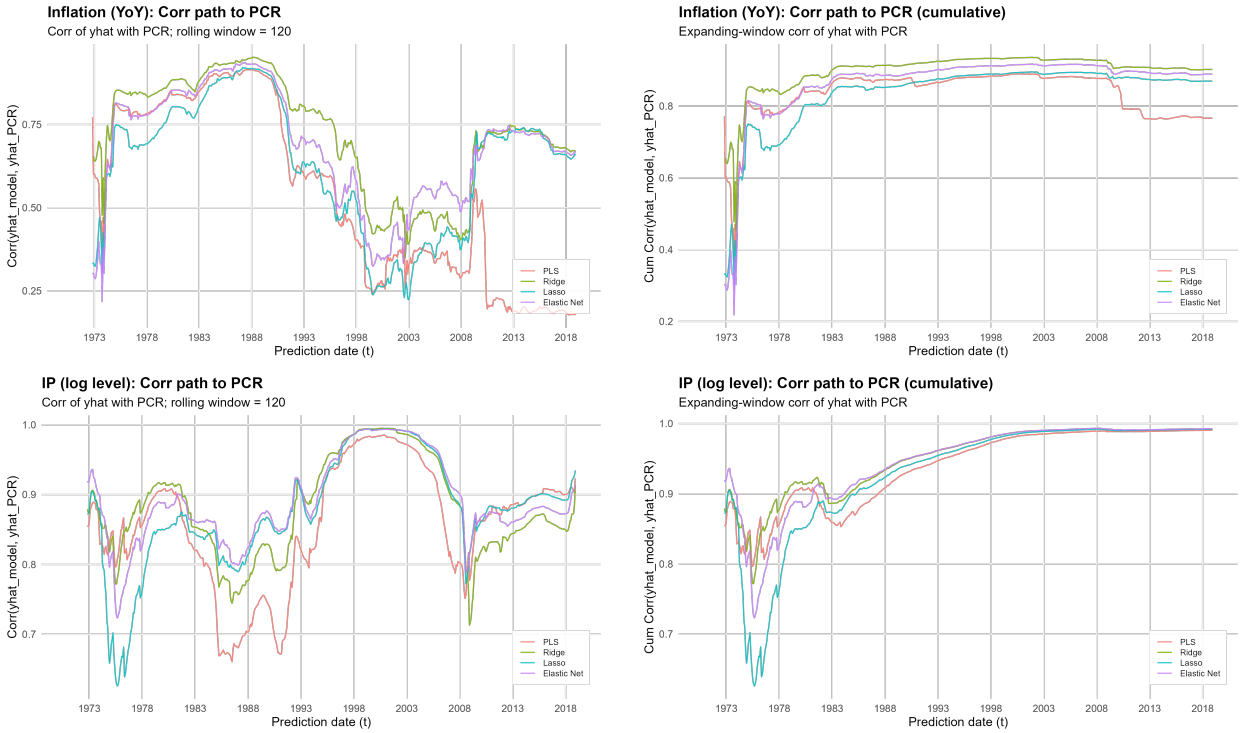
Table 3 reports forecast correlations between forecasts produced by alternative methods and those of PCR, to assess whether they extract similar predictive signals from the same information set. Over the full sample, forecast paths for both targets are highly correlated with PCR. Subsample results, however, reveal pronounced differences across variables. For industrial production, forecasts remain highly synchronized across all subsamples, indicating that regularization and dimension-reduction techniques rely on largely similar predictive combinations for this target. In contrast, for inflation, forecast alignment weakens markedly in later subsamples. Specifically, during 1985–2007, correlations between PLS and PCR and between Lasso and PCR fall to around 0.5; after 2008, forecast paths diverge further, with the PLS–PCR correlation declining to 0.20. To further examine the time variation in forecast alignment, we next report rolling and expanding correlation paths.

Figure 4 illustrates the time evolution of forecast path correlations. Overall, at short horizons (rolling correlations), forecast directions across methods exhibit episodic divergence; however, from a long-run cumulative perspective (expanding correlations), forecast paths tend to be pulled back toward similar directions. Ridge maintains the highest and most stable correlation with PCR across both targets, whereas PLS displays the most pronounced directional instability in inflation forecasting[2].

For industrial production (IP), forecast paths across methods remain highly synchronized at both short- and long-run horizons. Even during periods in which correlations decline—such as the mid-1980s Great Moderation period and around the 2008 financial crisis—the reductions occur in a largely synchronous

[2]Figure 14 in Appendix reports the distributions of cross-validated hyperparameters across methods and targets, providing additional descriptive evidence on how model complexity and regularization vary over time.

Figure 4: Correlations of Forecast Paths with principal component regression



Forecasts are generated using rolling estimation windows of fixed length $W = 120$ months. The distinction between rolling and expanding correlations reflects differences in how forecast similarity is evaluated.
Rolling correlations are computed as $\rho_t^{\text{roll}}(m, PCR) = \text{Corr}\Big(\{\widehat{y}^m_{s+h|s}\}_{s=t-W+1}^{t}, \{\widehat{y}^{\text{PCR}}_{s+h|s}\}_{s=t-W+1}^{t}\Big)$.
Expanding correlations are computed recursively as $\rho_t^{\text{exp}}(m, PCR) = \text{Corr}\Big(\{\widehat{y}^m_{s+h|s}\}_{s=1}^{t}, \{\widehat{y}^{\text{PCR}}_{s+h|s}\}_{s=1}^{t}\Big)$.

manner across methods. In expanding correlations, alignment increases almost monotonically and converges rapidly toward unity. This pattern indicates that differences across methods primarily reflect short-run adjustments in directional weights, while long-run predictive signals remain common.

In contrast, for inflation (INF), forecast paths across methods exhibit substantial and persistent divergence in the middle and later parts of the sample. This divergence is evident not only in the pronounced volatility of rolling correlations but also in the failure of expanding correlations to return to high levels over time. In particular, the correlation dynamics of PLS differ from those of the three shrinkage-based methods, with less synchronized movements and greater fluctuations, indicating that forecast directions do not naturally converge across methods as the sample expands

The empirical results are consistent with the simulation analysis. Ridge and PCR exhibit the highest and most stable forecast path correlations across both targets, which is consistent with their mechanism of continuously shrinking coefficients while averaging information over the dominant eigenspace. Elastic Net performs similarly, reflecting its endogenous tendency to tilt toward ridge-type regularization in environments with high collinearity. By contrast, Lasso relies on discrete selection among highly correlated predictors, resulting in forecast path correlations that are lower than those of Ridge and Elastic Net and in coefficient choices that vary more across samples. PLS, which ranks and truncates candidate supervised directions, is the most sensitive to the predictability of the target and the underlying covariance structure, and therefore displays the most pronounced instability in its correlation with PCR in inflation forecasting.

5 Conclusion

The main finding of this paper is that whether different dimension-reduction and regularization methods generate highly correlated forecast paths in high-dimensional predictive regressions depends primarily on how predictive information is organized in the covariance structure of the predictors, rather than on whether a method is classified as compression- or shrinkage-based. When predictive signals are concentrated in the dominant eigenspace of the predictor covariance matrix, differences in tuning choices, finite-sample performance, and estimation paths do not prevent alternative methods from recovering highly similar predictive directions as the cross-sectional and time dimensions grow jointly. By contrast, when the predictive covariance structure is unstable or the predictive signal is weakly identified, divergences in forecast paths can persist in finite samples.

Ridge and principal component regression exhibit the highest and most stable forecast path correlations in both simulations and empirical applications, with correlations in the simulations increasing systematically and approaching unity as (n, T) increase. This result reflects the asymptotically equivalent ways in which the two methods exploit the dominant eigenspace: PCR applies a discrete truncation of the feature space, whereas Ridge uses continuous shrinkage to average information within the same space. Elastic Net produces forecast paths that closely track those of Ridge and, in the simulations, converges toward Ridge as dimensionality increases, reflecting its endogenous tendency to tilt toward ridge-type regularization in highly collinear environments. In contrast, Lasso relies on discrete selection among highly correlated predictors, leading to lower forecast path correlations with PCR in finite samples, although alignment improves as the predictive subspace becomes more precisely identified.

The behavior of partial least squares is most sensitive to the structure of predictive covariance. In the simulations, PLS aligns closely with PCR when the predictive covariance exhibits a clear and stable dominant direction, but convergence is slower and forecast paths can deviate in environments where weak or heterogeneous noise factors contaminate the leading eigenspace. The empirical results mirror these asymptotic patterns: for industrial production, forecast paths across methods rapidly converge as the sample expands, whereas for inflation, forecast path divergence persists over time, consistent with the absence of a stable dominant predictive direction in the underlying covariance structure.

References

Bai, J. (2003). Inferential theory for principal components analysis in large factor models. *Econometrica*, 71(1):135–173.

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

De Mol, C., Giannone, D., and Reichlin, L. (2024). The asymptotic equivalence of ridge and principal component regression with many predictors. *Econometrics and Statistics*.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Appendix

Appendix A. Technical Details

This appendix collects two technical facts used to interpret why high-dimensional compression and shrinkage forecasts often align in large panels: (i) linear ℓ_2-shrinkage acts as a *spectral filter* in the principal-component basis; (ii) under spectral separation, the leading eigenspace is consistently estimated, yielding stable projections onto the dominant subspace.

A.1 Spectral-filter representation of Ridge and PCR

Let $X_t \in \mathbb{R}^n$ denote predictors and let y_{t+h} be the target. On a training sample of size T_{tr}, stack predictors in the $T_{\text{tr}} \times n$ matrix X and targets in $y \in \mathbb{R}^{T_{\text{tr}}}$. Define the sample covariance

$$\widehat{\Sigma}_X \;=\; \frac{1}{T_{\text{tr}}} X'X \;=\; \widehat{V}\,\widehat{\Lambda}\,\widehat{V}', \qquad \widehat{\Lambda} = \operatorname{diag}(\widehat{\mu}_1, \ldots, \widehat{\mu}_n),\ \widehat{\mu}_1 \geq \cdots \geq \widehat{\mu}_n \geq 0,$$

and let $\widehat{Z} := X\widehat{V}$ denote the matrix of sample principal-component scores (columns $\widehat{z}_j = X\widehat{v}_j$).

Lemma A.1 (Ridge as PC-wise shrinkage). Consider ridge regression on the training sample,

$$\widehat{\beta}^{\text{Ridge}}(\lambda) = \arg\min_{\beta \in \mathbb{R}^n} \; \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \qquad \lambda > 0.$$

Then the in-sample fitted values admit the spectral-filter representation

$$\widehat{y}^{\text{Ridge}}(\lambda) := X\widehat{\beta}^{\text{Ridge}}(\lambda) = \sum_{j=1}^{n} g_\lambda(\widehat{\mu}_j)\, \widehat{P}_j\, y, \qquad g_\lambda(\widehat{\mu}) := \frac{\widehat{\mu}}{\widehat{\mu} + \lambda}, \quad \widehat{P}_j := \frac{1}{T_{\text{tr}}} \widehat{z}_j \widehat{z}_j'.$$

Equivalently,

$$\widehat{y}^{\text{Ridge}}(\lambda) = \frac{1}{T_{\text{tr}}} \widehat{Z}\, \operatorname{diag}\left(\frac{T_{\text{tr}}\widehat{\mu}_j}{T_{\text{tr}}\widehat{\mu}_j + \lambda} \right)_{j=1}^{n} \widehat{Z}' y.$$

Proof. The ridge normal equations give $\widehat{\beta}^{\text{Ridge}}(\lambda) = (X'X + \lambda I_n)^{-1} X'y$. Using $X'X = T_{\text{tr}} \widehat{V}\widehat{\Lambda}\widehat{V}'$,

$$X(X'X + \lambda I_n)^{-1} X' = X\widehat{V}\,(T_{\text{tr}}\widehat{\Lambda} + \lambda I_n)^{-1} \widehat{V}' X' = \widehat{Z}\,(T_{\text{tr}}\widehat{\Lambda} + \lambda I_n)^{-1} \widehat{Z}'.$$

Since $\widehat{Z}' (= \widehat{V}'X')$ and $\widehat{Z}'\widehat{Z} = T_{\text{tr}}^2 \widehat{\Lambda}$, this yields the claimed diagonal weighting in the PC-score basis. □

Corollary A.1 (PCR as hard spectral truncation). Let $\widehat{V}_r = (\widehat{v}_1, \ldots, \widehat{v}_r)$ and $\widehat{Z}_r = X\widehat{V}_r$. The principal component regression (PCR) fitted values using r components satisfy

$$\widehat{y}^{\text{PCR}}(r) = \widehat{Z}_r (\widehat{Z}_r'\widehat{Z}_r)^{-1} \widehat{Z}_r' y = \sum_{j=1}^{r} \widehat{P}_j\, y.$$

Thus PCR corresponds to the *hard-threshold* spectral filter $g^{\text{PCR}}(\widehat{\mu}_j) = \mathbf{1}\{j \leq r\}$, while ridge corresponds to the *soft* spectral filter $g_\lambda(\widehat{\mu}_j) = \widehat{\mu}_j / (\widehat{\mu}_j + \lambda)$.

Interpretation. Both procedures form fitted values by projecting y onto PC-score directions $\{\widehat{z}_j\}$, but they treat weak directions differently: PCR discards them entirely, ridge shrinks them continuously. This difference is precisely what becomes less consequential when the predictive signal is concentrated in a dominant eigenspace and weak directions carry limited predictive content.

A.2 Eigenspace stability under spectral separation

Let $\Sigma_X = \mathbb{E}(X_t X_t')$ denote the (population) predictor covariance and $\Sigma_X = V\Lambda V'$ its eigendecomposition, with eigenvalues $\mu_1 \geq \cdots \geq \mu_n \geq 0$ and eigenvectors $V = (v_1, \ldots, v_n)$. Let $V_r = (v_1, \ldots, v_r)$ and define the population projector onto the dominant eigenspace $P_r := V_r V_r'$. Similarly, define $\widehat{P}_r := \widehat{V}_r \widehat{V}_r'$ based on $\widehat{\Sigma}_X$.

Lemma A.2 (Davis–Kahan bound for projectors). Let the *eigengap* at r be

$$\mathrm{gap}_r \;:=\; \min\{\mu_r - \mu_{r+1},\ \mu_{r-1} - \mu_r\},$$

with the convention that $\mu_0 = +\infty$ and $\mu_{n+1} = -\infty$. If $\mathrm{gap}_r > 0$, then

$$\|\widehat{P}_r - P_r\|_{\mathrm{op}} \;\leq\; \frac{2\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}}}{\mathrm{gap}_r}.$$

Proof. This is a standard Davis–Kahan (or Wedin) sin-Θ perturbation inequality for invariant subspaces, specialized to orthogonal projectors. □

A strong-factor implication (AFS-type scaling). In the strong-factor (AFS) environments used in the simulations, the leading eigenvalue diverges linearly while the remaining eigenvalues are bounded:

$$\mu_1(\Sigma_X) \asymp n, \qquad \mu_2(\Sigma_X) = O(1).$$

Hence $\mathrm{gap}_1 = \mu_1 - \mu_2 \asymp n$, and Lemma A.2 yields

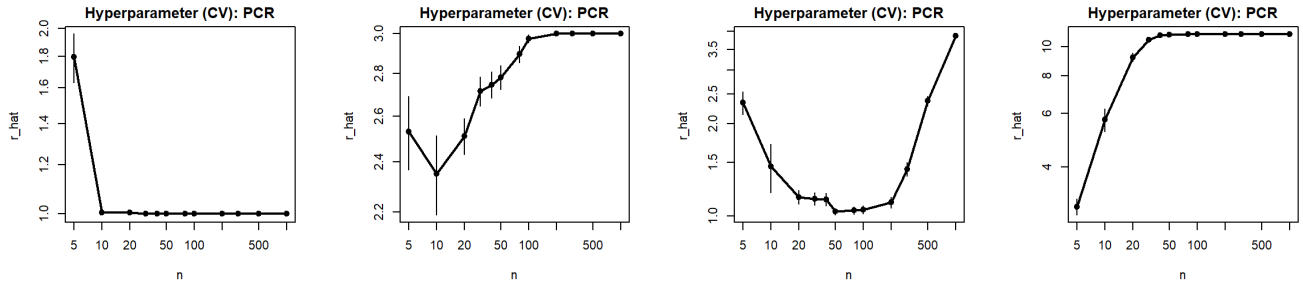
$$\|\widehat{P}_1 - P_1\|_{\mathrm{op}} \;\leq\; \frac{2\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}}}{c\,n} \qquad \text{for some } c > 0.$$

Therefore, any bound implying $\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}} = o_p(n)$ is sufficient for consistency of the dominant projector $\widehat{P}_1$. More generally, when the dominant cluster is separated from the remainder (as in AFS and in milder GFS regimes), $\widehat{P}_r$ is stable, which supports the empirical finding that distinct procedures that effectively emphasize the dominant eigenspace generate highly aligned forecast paths.

Remark (Why this matters for forecast alignment). Combining Lemma A.1 and Lemma A.2: if predictive content is concentrated in the subspace P_r and $\widehat{P}_r$ is stable, then (i) PCR, which hard-truncates to $\widehat{P}_r$, and (ii) ridge, which heavily weights the same leading directions (large $\widehat{\mu}_j$) while shrinking the rest, will tend to produce highly correlated fitted values and hence similar forecast paths out of sample.

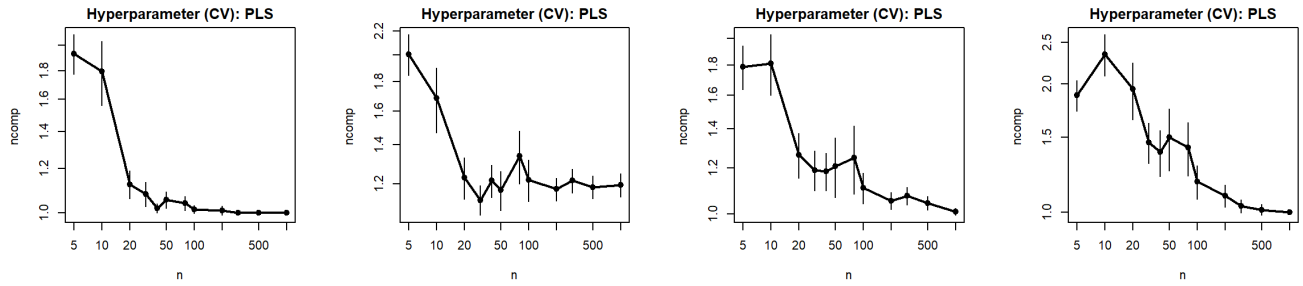
Appendix B. Simulation Details

Figure 5: PCR: number of principal components



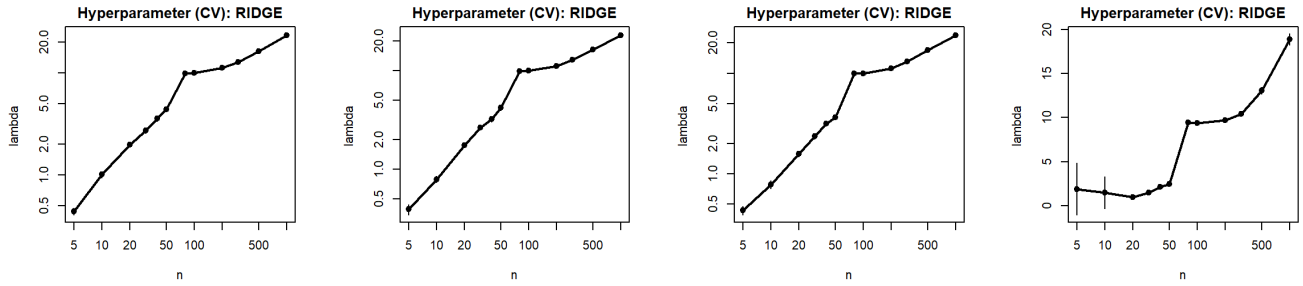
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 6: PLS: number of latent components



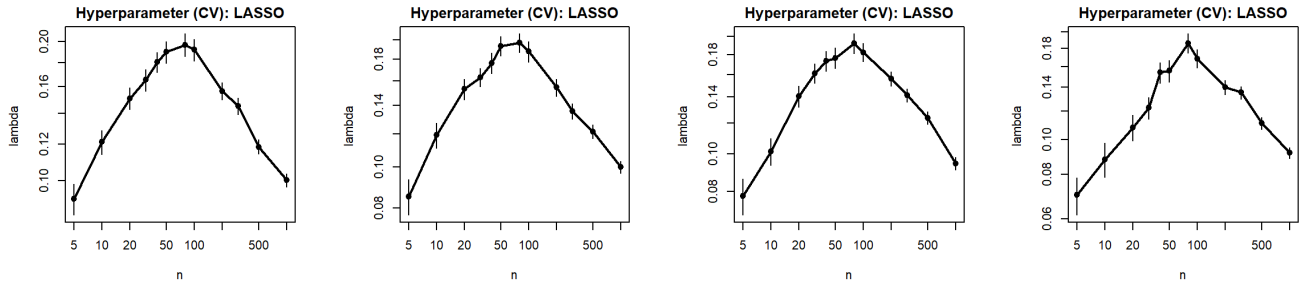
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 7: Ridge: regularization parameter λ



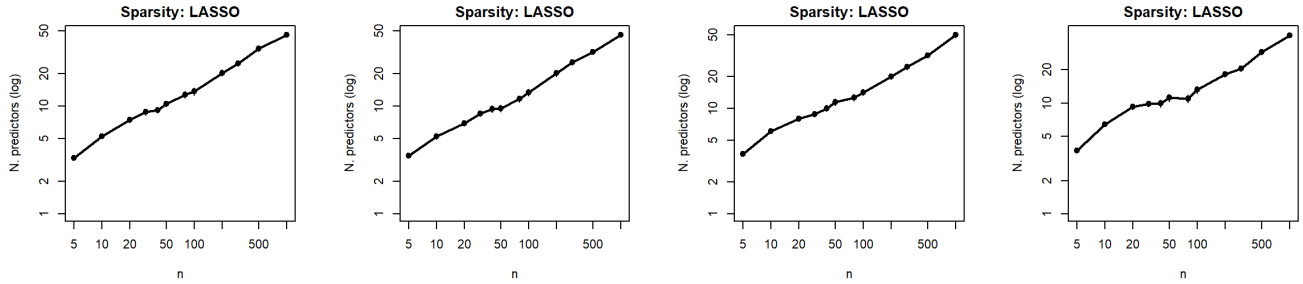
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 8: Lasso: regularization parameter λ



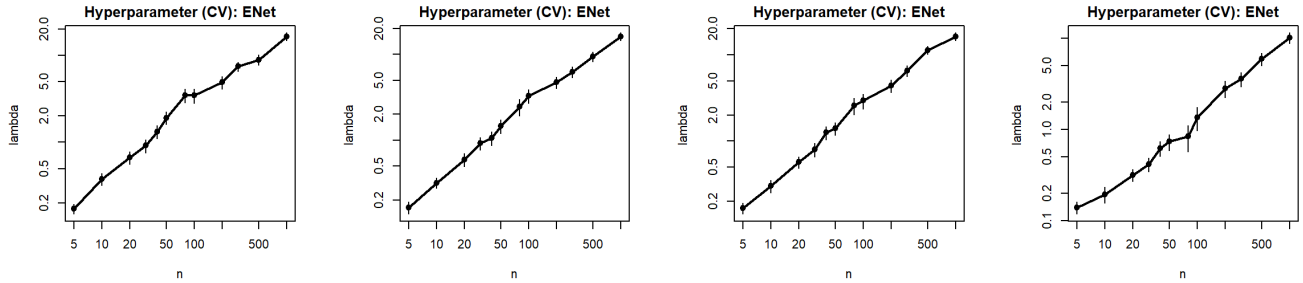
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 9: Lasso: number of selected predictors



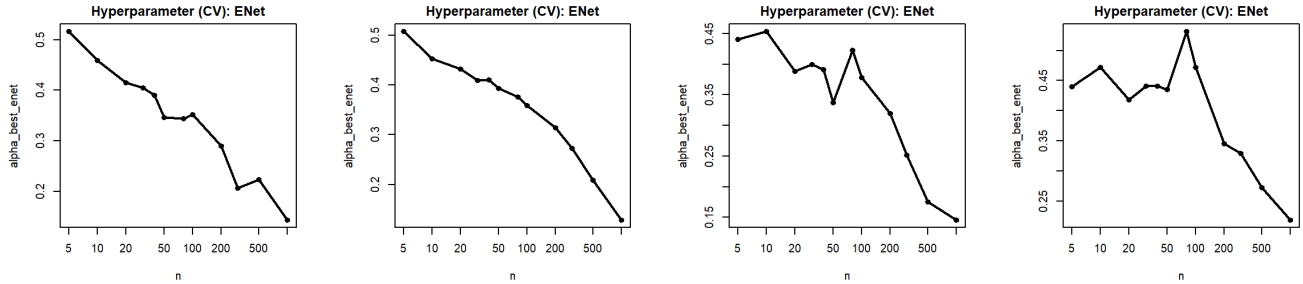
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 10: Elastic Net: regularization parameter λ



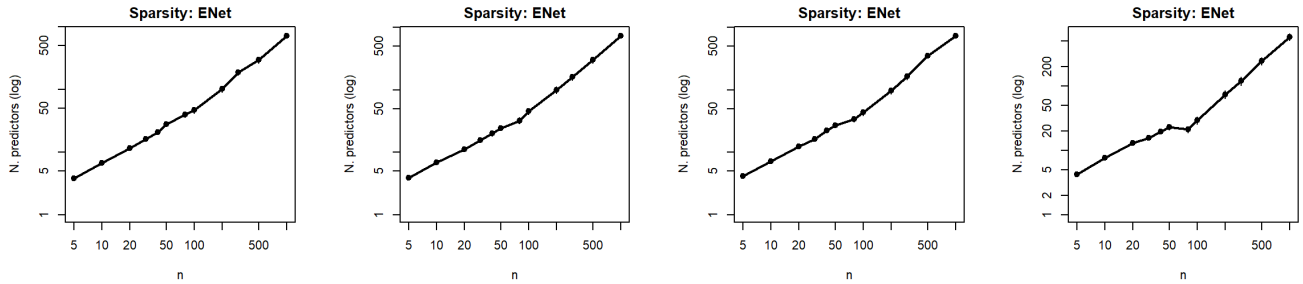
(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

Figure 11: Elastic Net: mixing parameter α



(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

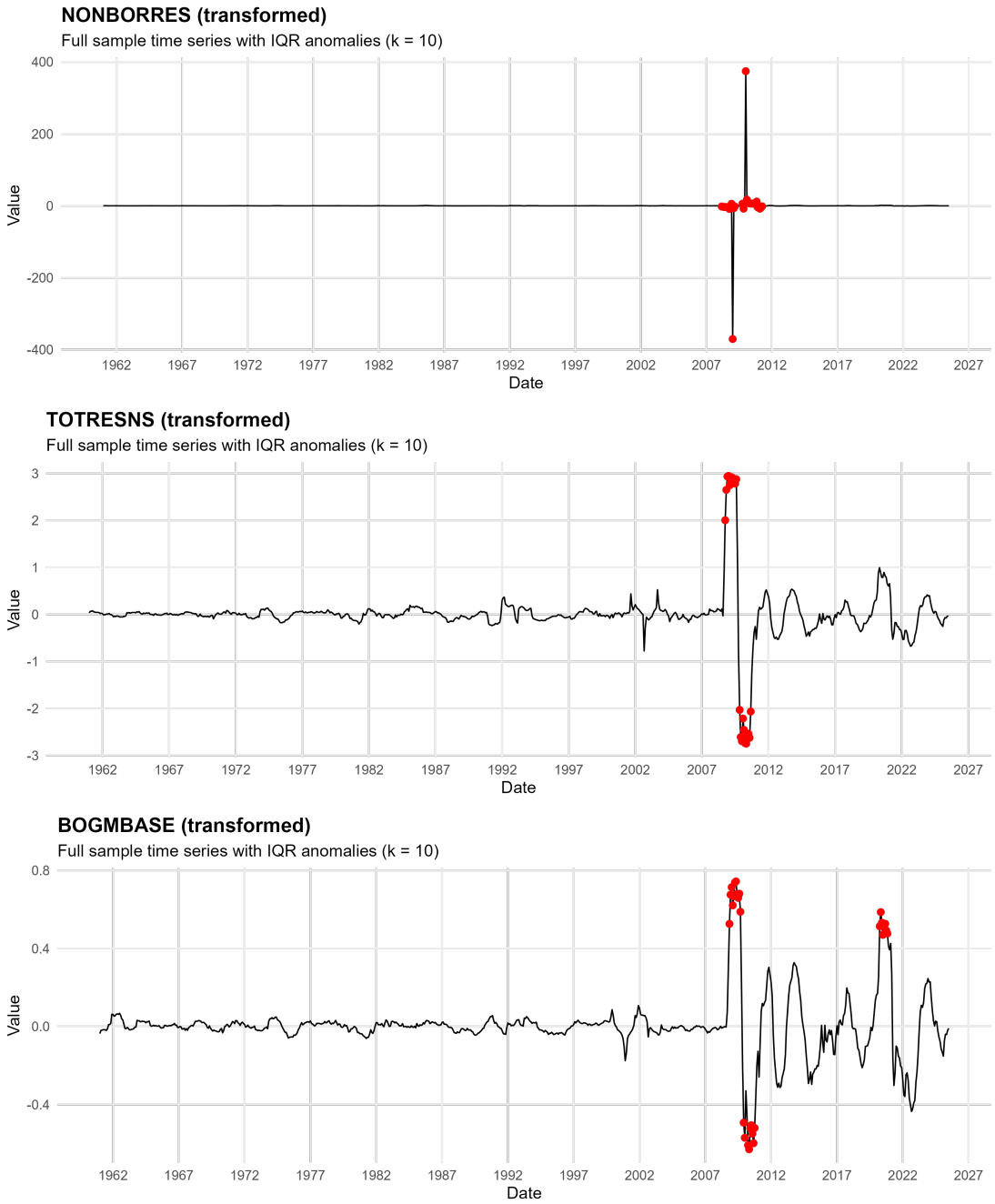
Figure 12: Elastic Net: number of selected predictors



(a) AFS (b) WFS (c) GFS–mild (d) GFS–hard

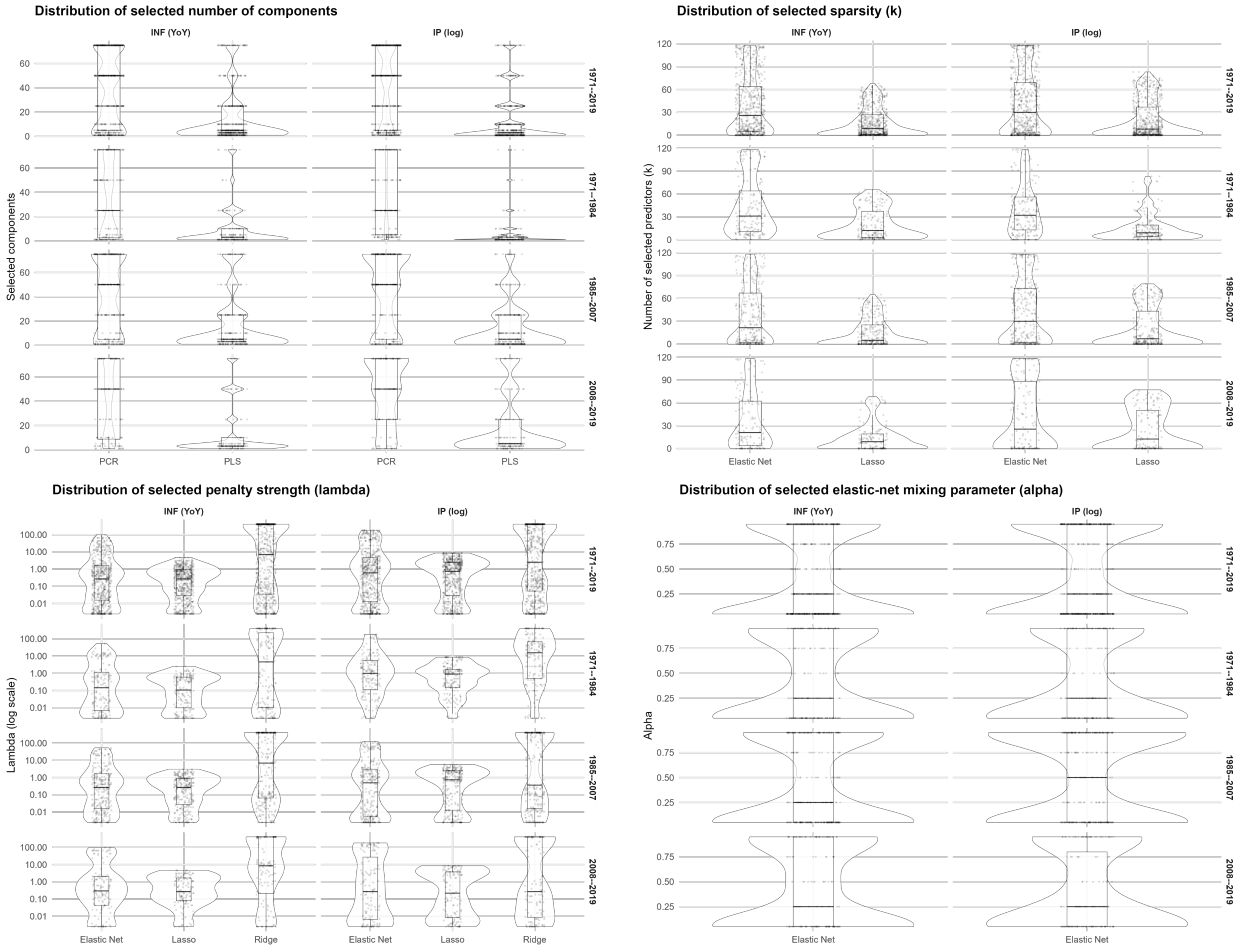
Appendix C. Empirical Details

Figure 13: IQR-Based Anomaly Detection for Reserve Aggregate



Red dots indicate observations flagged as extreme based on an interquartile range (IQR) rule with threshold $k = 10$. In the sample, NONBORRES (non-borrowed reserves) is excluded. As a residual series defined as total reserves minus borrowed reserves, NONBORRES exhibits extremely large outliers during 2008–2010, likely driven by changes in borrowing facilities and statistical composition rather than underlying economic variation. These observations are not comparable in scale to the rest of the sample. By contrast, TOTRESNS (total reserves) and BOGMBASE (monetary base) directly reflect the implementation of monetary policy in response to the financial crisis. Although they display substantial fluctuations, these movements remain limited in magnitude and economically interpretable, and are therefore retained in the baseline analysis.

Figure 14: Distribution of selected hyperparameters by cross-validation



This figure reports the empirical distribution of hyperparameters selected by time-series cross-validation across rolling forecasting windows. Panel (a) shows the number of components selected by PCR and PLS, Panel (b) reports the sparsity level measured by the number of nonzero coefficients selected by Lasso and Elastic Net, Panel (c) displays the regularization parameter λ (on a $\log_{10}$ scale) for Ridge, Lasso, and Elastic Net, and Panel (d) reports the Elastic Net mixing parameter α. Each row corresponds to a sample period (full sample and subsamples), and columns distinguish the two target variables. Violin plots depict kernel density estimates, boxplots indicate the interquartile range and median, and dots represent window-specific hyperparameter choices.