VERSOR
INVESTMENTS

Contextual Alpha: Emphasizing Forecasts Where They Work Best

Ludger Hentschel

July 22, 2025

Abstract
We propose a context-aware framework for turning raw predictive signals into stock-level alphas that identify and exploit cross-sectional heterogeneity. A panel regression decomposes the signal's slope, its local information coefficient (IC), into additive contributions from economically motivated, potentially overlapping groups.

We estimate group-level IC components in a panel regression with generalized Ridge (empirical-Bayes) shrinkage. The shrinkage pulls the global intercept toward the pooled IC, penalizes imprecise local effects in proportion to their estimated covariance, and leaves time-varying fixed effects unpenalized so that each group is properly centered cross-sectionally.

After estimation we can solve a covariance-weighted quadratic program that moves the coefficients statistically as little as necessary to ensure that every stock's implied IC is non-negative, reflecting the economic prior that the signal's direction should not flip.

Multiplying the stock-specific ICs by the raw signal values produces group-aware, contextual alphas that amplify the signal where evidence is strong and mute it where it is weak. These alphas can be used in portfolio construction or can be added to a fundamental multi-factor risk model to derive orthogonalized factor exposures.

The procedure generalises seamlessly to many signals, each with its own contextual hierarchy, and accommodates overlapping group definitions without proliferating thousands of dummy-interaction factors.

www.versorinvest.com   lhentschel@versorinvest.com
1120 Avenue of the Americas, 15th Floor, New York, NY 10036

Contents

1 Introduction

A core challenge in quantitative investing is measuring and improving the predictive performance of trading signals. A common approach estimates the average signal strength, typically measured by the information coefficient (IC), across the full investable universe. This assumes uniform signal effectiveness across all stocks, despite substantial evidence that even well-established signals, such as short-term reversal or post-earnings drift, perform differently depending on sector, industry, firm size, and proximity to corporate events.

I introduce a contextual alpha framework to address this heterogeneity. The proposed framework allows the information coefficient (IC) of a signal to vary systematically across pre-defined stock groups. By representing ICs as additive functions of group-level categorical features, such as sector, industry, size decile, and event time, we model how predictive strength depends on the economic and informational environment surrounding each stock. This approach is intuitive, data-driven, and practically useful for refining signals and improving portfolio performance.

Our main tool is a group-structured regression in which the slope of the signal varies by group membership. The slope estimates correspond to local ICs, rank correlations between the signal and future returns, within each subgroup. We can estimate this model using regularized regression, which balances flexibility with estimation stability by shrinking weak or noisy group effects toward zero.

To improve interpretability and enforce economic discipline, we can post-process the estimated group ICs to ensure they are non-negative across all observed groups. This adjustment is posed as a constrained quadratic optimization problem that balances statistical precision with prior beliefs.

The estimated group-specific ICs are then used to compute expected returns, which serve as inputs to portfolio construction. This yields group-aware return forecasts that allocate more weight to signal components with higher empirical support, while de-emphasizing groups where the signal is weak or unreliable. The resulting portfolios are interpretable, economically consistent, and empirically robust.

The proposed framework accommodates overlapping group structures and extends naturally to multi-signal settings. It provides an interpretable decomposition of signal performance, allowing managers to diagnose, refine, and implement signals with greater precision.

The focus of improving a single signal with conditioning information and machine learning methods differs from the more common idea of using machine learning methods to combine a large number of predefined signals. See Gu, Kelly, and Xiu (2020), Chen, Pelger, and Zhu (2024), or Li, Rossi, Yan, and Zheng (2025), for example. Müller and Schmickler (2025) follow an intermediate approach by investigating all pairwise interactions between signal candidates. When combining multiple signals, it is possible to interpret some of the signals as conditioning information. But the forecast does not separate signals from conditioning information. A possible exception is macroeconomic data that is generally interpreted as conditioning information for the signals.

Howard (2024) investigates separate nonlinear factor models across 3 size groups and concludes that such models outperform a single, homogenous model. The nonlinearity in the factor models may condition some signals on supplied characteristics other than size. Unfortunately, such conditioning is hidden inside the models. Also, there is no provision of shrinkage, which becomes important as the number of groups grows, or allowance for overlapping groups.

The remainder of the paper proceeds as follows: Section 2 presents the group-structured regression and interprets its coefficients as local ICs. Section 3 discusses estimation and regularization. Section 4 derives expected returns from the estimated ICs and explains how to use them in portfolio construction. Section 5 discusses how to use contextual alphas in a multifactor risk model and section 6 concludes.

2 Group-Structured Regressions for ICs

In this framework, we aim to quantify how the predictive effectiveness of a trading signal varies across groups of stocks defined by categorical characteristics such as sector, industry, size decile, and event time (e.g., proximity to earnings announcements). Based on empirical findings, we can amplify the signal where it works best and mute the signal where it works least.

This is a form of tuning for individual signals that permits more flexibility than standard signal constructions while encouraging parsimony. To improve the odds that the additional flexibility is useful, we choose specific stock characteristics that may affect signal efficacy. Because we tune individual signals, these characteristics may vary across signals. This differs from Freyberger, Neuhierl, and Weber (2020), for example, who fit univariate nonparametric functions to signals, without conditioning on stock characteristics, thereby treating all firms homogeneously.

We start with standard ingredients for predictive regressions. Let $r_{i,t+1}$ denote the within-industry rank of future returns for stock i at time $t+1$, and let $s_{i,t}$ denote the within-industry rank of the current signal value for the same stock on a preceding date. For returns and signals, we center and standardize the ranks, so that they have mean zero and standard deviation of one within each group.

The mainstream approach is to model the return rank across the entire investable universe as a single linear function of the signal

$$r_{i,t+1} = \beta s_{i,t} + \epsilon_{i,t}. \tag{1}$$

If the ranks of returns and signals are centered around zero, the intercept in this regression is equal to zero. Due to the centered and standardized ranking of the returns and signals, we can interpret β as the rank correlation between future returns and current signals. In stock return prediction, we commonly refer to this rank correlation as the information coefficient.

Following Fama and MacBeth (1973), it is common to run a sequence of cross-sectional regressions and then average betas over time. Running a panel regression with time-invariant betas essentially reproduces this average beta. In order to accommodate a substantial number of parameters that are meant to be stable over time, I will focus on panel regressions.

The signals $s_{i,t}$ can be Fama and French (1992) size or value factors, Jegadeesh and Titman (1993) momentum factors, or any of the large number of equity factors listed in Hou, Xue, and Zhang (2018), for example. Many proprietary signals appear to take this form as well. Of course, this structure is not confined to equity investing, even though it seems most prominent there.

2.1 Group-Structured Regression

A natural generalization models the return rank as a linear function of the signal, where the IC varies additively across groups

$$\begin{aligned} r_{i,t+1} = \Bigg(& \beta_0 + \sum_{g \neq g_0} \beta_g^{\text{sec}} D_{ig}^{\text{sec}} + \sum_{h \neq h_{g(i)}} \beta_h^{\text{ind}} D_{ih}^{\text{ind}} \\ & + \sum_{d \neq d_0} \beta_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \beta_e^{\text{event}} D_{ie}^{\text{event}} \Bigg) s_{i,t} + \varepsilon_{i,t}. \end{aligned} \tag{2}$$

Here

- β_0 captures the baseline effectiveness of the signal across the universe.

- D_{ig}^{sec} is a dummy equal to 1 if stock i belongs to sector g; one sector (indexed by g_0) is omitted for identifiability.
- D_{ih}^{ind} is a dummy for industry h within sector $g(i)$; one industry is omitted in each sector.
- D_{id}^{size} is a dummy for size group d, omitting one reference group (e.g., the largest stocks).
- D_{ie}^{event} indicates event-time classification (e.g., pre- or post-earnings), omitting a reference group such as "normal" periods.

These dummy categories illustrate the kinds of structures we can model. Although they seem economically reasonable, they are not meant to be the best groupings. Different signals may benefit from different groupings. This signal-specific modeling of groups is a potential benefit of this approach compared to using the same groups for all signals.

The dummy variables provide conditioning information for the signal but we assume that the group information does not have predictive value on its own.

This formulation ensures identifiability: each group-specific coefficient β^j is interpreted as a deviation from its respective reference group. For example, β_1^{size} measures the difference in signal effectiveness for stocks in size group 1 relative to the omitted group d_0.

This is a richly parameterized functional form that allows different ICs for each labeled group but reduces the potential number of parameters by not interacting the groups with each other.[1]

Especially in the context of noisy return predictions, the additional flexibility and parameters of this approach raise concerns about estimation noise. To balance flexibility with a reduction of noise, we can apply regularization to the coefficients. Regularization shrinks noisy or weak deviations toward zero, effectively collapsing the model to a sparse structure where the signal is assumed equally effective across many groups. In an extreme case, only β_0 remains, corresponding to a uniform signal effect.

Because both the dependent variable $r_{i,t+1}$ and the signal $s_{i,t}$ are within-industry ranks, the slope coefficients can be interpreted as Spearman rank correlations, or ICs. The baseline β_0 reflects the signal's effectiveness in the reference group, and each group-specific coefficient (e.g., β_g^{sec}) captures deviations from that baseline. The sum $\beta_0 + \beta_g^j$ gives the effective IC for group g.

[1] Mechanically, adding interactive terms would be a straightforward extension of this framework. But it quickly introduces a very large number of parameters. I do not pursue this analysis here.

Although the model no longer yields a single global IC, it enables a granular decomposition across sectors, industries, and other dimensions. A weighted average of group-level slopes still approximates the overall IC

$$\widehat{\beta}_{\text{pooled}} = \frac{\sum_g \sum_{i \in g} r_{i,t+1} s_{i,t}}{\sum_g \sum_{i \in g} s_{i,t}^2} = \sum_g \left(\frac{V_g}{\sum_g V_g} \right) \widehat{\beta}_g, \tag{3}$$

where $V_g = \sum_{i \in g} s_{i,t}^2$ is the signal variance in group g. Because we standardize signals and returns, signal variances are equal across groups and the pooled estimate simplifies to the weighted average of the group-specific betas

$$\widehat{\beta}_{\text{pooled}} = \sum_g \left(\frac{n_g}{N} \right) \widehat{\beta}_g. \tag{4}$$

In this structure, we modulate the strength of signal across groups, depending on the signal's effectiveness in the groups. The signal can be larger in some groups than in others. In most portfolio constructions, this will create larger gross exposures in the groups with larger signals. We maintain the linear relation between returns and signal but allow the slope coefficient to vary across groups. Hanauer, Soebhag, Stam, and Hoogteijling (2025) investigate separate nonlinear functions linking returns and signals in each sector but conclude that a common function across all sectors provides better out-of-sample predictions.

2.2 Time-Varying Fixed Effects

The regression models in equation (1) and equation (2) have zero intercepts when returns and signals are centered within each group. In equation (2) that condition may not hold for all groups. To account for potential non-zero means in signal values or returns within some groups, we can add fixed effects for each of these groups. To properly center the groups each period, we have to introduce separate group dummies for each period. This ensures that each group is properly centered in each cross-section, preventing spurious correlations from driving estimates of signal effectiveness.

The original regression model augmented with time-varying group dummies is

$$\begin{aligned} r_{i,t+1} = \Bigg(& \beta_0 + \sum_{g \neq g_0} \beta_g^{\text{sec}} D_{ig}^{\text{sec}} + \sum_{h \neq h_{g(i)}} \beta_h^{\text{ind}} D_{ih}^{\text{ind}} \\ & + \sum_{d \neq d_0} \beta_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \beta_e^{\text{event}} D_{ie}^{\text{event}} \Bigg) s_{it} \end{aligned}$$

$$+\sum_t \sum_d \gamma^{\text{size}}_{d,t} D^{\text{size}}_{id,t} + \sum_t \sum_e \gamma^{\text{event}}_{e,t} D^{\text{event}}_{ie,t} + \varepsilon_{it}. \tag{5}$$

Here, the slope coefficients β are time-invariant and measure how signal effectiveness varies across group memberships. The intercept terms $\gamma_{d,t}$, $\gamma_{e,t}$, are time-varying fixed effects that control for non-zero group means in each period. The regressions intentionally omit time-varying fixed effects for sectors and industries under the assumption that the signals and returns are centered within sectors and industries. When that is not the case, we should include dummies for these groups as well.

We know that these fixed effects are zero for groups where we have centered the signals and returns and we can exclude them there. In our example we have assumed that signals and returns are centered within each industry and therefore within each sector. Hence, we can exclude the fixed effects γ_g and γ_h for sector and industries, respectively.

By separating intercepts from slopes, the model cleanly decomposes average group-level differences in returns (via γ^j), and group-varying signal effectiveness (via β^j). This separation ensures that the estimated ICs β^j reflect true differences in signal performance, not artifacts of group-level return shifts or uncentered signal distributions.

This is particularly relevant when the group means of the returns or signals are not zero. For example, if small-cap stocks have consistently higher average return ranks, the fixed effects $\gamma^{\text{size}}_{d,t}$ will absorb this shift. Or, if signals are not mean-zero within each group, the group-specific intercepts help isolate the variation in returns explained by the signal.

2.3 Enforcing Positivity

Campbell and Thompson (2008) show that enforcing sign constraints on return forecasts can improve predictive accuracy.[2] Here, I apply a related idea: if we believe a signal should yield non-negative ICs across all groups, we can adjust the estimated group coefficients to enforce this economic prior. Note carefully that this permits zero ICs, but not negative ICs.

Let $\widehat{\boldsymbol{\beta}}$ denote the vector of estimated slope coefficients and $\widetilde{\boldsymbol{\Omega}}$ the covariance matrix of the estimated slope coefficients.[3] We seek an adjusted coefficient vector $\widetilde{\boldsymbol{\beta}}$ that is statistically close to $\widehat{\boldsymbol{\beta}}$ but ensures non-negative implied ICs for all observed group combinations.

[2] Jagannathan and Ma (2003) demonstrate a similar effect in portfolio construction and risk forecasting. Gu, Kelly, and Xiu (2020) and Chen, Pelger, and Zhu (2024) show that absence of arbitrage constraints improve predictions in their models.

[3] I will discuss later that $\widetilde{\boldsymbol{\Omega}}$ is a shrunk version of the covariance matrix $\boldsymbol{\Omega}$.

The quadratic program

$$\min_{\widetilde{\boldsymbol{\beta}}} \quad (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})' \widetilde{\boldsymbol{\Omega}}^{-1} (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \tag{6}$$

subject to

$$\boldsymbol{d}_j' \widetilde{\boldsymbol{\beta}} \geq 0 \quad \text{for all group combinations } j, \tag{7}$$

delivers such an adjusted estimate of the slope coefficients. The quadratic form punishes large changes in precisely estimated coefficients.

Here, $\boldsymbol{d}_j$ is a dummy vector corresponding to group combination j. In the regression, each row of the design matrix has the form $\boldsymbol{x}_{it} s_{it}$, where $\boldsymbol{x}_{it}$ contains the group dummies and intercept. To identify the constraint vectors $\boldsymbol{d}_j$, we take the unique rows of $\boldsymbol{x}_{it}$ (across all observations) and treat their transposes as the set of dummy vectors $\boldsymbol{d}_j$.

This adjustment has at least two desirable properties. First, it enforces our prior belief that the signal is non-negative everywhere. Second, it preserves group coefficients with strong empirical support.

3 Estimation

The slope coefficients in equation (2) intentionally do not carry time subscripts. While the IC and its components may change over time, the main objective is to model persistent differences across groups of stocks. The most reliable method for estimating these stable coefficients is in a panel regression. We can stack many periods into a panel and then estimate the slope coefficients.

To accommodate slow changes in the coefficients, we can certainly run the panel regression for shorter or rolling windows. Alternatively, we can give more weight to recent periods and less weight to distant periods.

In principle, we can apply any regularized regression framework in estimating the model. The leading candidates employ different combinations of ℓ_1 and ℓ_2 regularization. The ℓ_1 regularization in the Tibshirani (1996) Lasso regression strongly favors parsimony by setting some of the coefficients to 0. The ℓ_2 regularization in the Hoerl and Kennard (1970) Ridge regression shrinks estimates toward 0 but does not collapse the estimates there. The combination of ℓ_1 and ℓ_2 regularization of the Zou and Hastie (2005) Elastic Net regression allows for elements of both.

Shen and Xiu (2024) and Kozak, Nagel, and Santosh (2020) argue that ℓ_2 regularization is better able to learn a large collection of weak signals than ℓ_1

regularization. This seems intuitive, but it is not clear that their logic applies hear. In this setting, we have selected a signal. The main effort now is to refine this signal. It is not clear that a large number of small refinements, and the associated estimation noise, produce better forecasts than a small number of larger adjustments. This is an empirical question and the answer may depend on the signal.

The main challenge in estimating the regularized regression in equation (2) is to apply a key insight from shrinkage estimates: The amount of shrinkage should depend on the precision of the empirical estimate. Since the groups we defined are likely to contain different numbers of observations, the corresponding coefficient estimates are likely to have different precisions. The uniform shrinkage in standard Lasso, Ridge, and Elastic Net regressions is not a great match for this problem.

Hoerl and Kennard (1970) also defined a generalized Ridge regression that allows for different penalties for each coefficient. If the groups don't overlap or the coefficient estimates are approximately uncorrelated, penalties in proportion to the inverse of the squared standard errors are attractive. When the groups overlap or the coefficient estimates are correlated, using the full covariance matrix of the estimates as a shrinkage penalty becomes attractive.

Similar, generalized versions of the Lasso and Elastic Net regressions are feasible. Tibshirani and Taylor (2011) discuss generalized Lasso regressions. Hellum, Pedersen, and Rønn-Nielsen (2024) discuss generalized Elastic Net estimates in the context of global multi-factor models. I will focus on generalized Ridge regression in the remainder. A key advantage of this choice is that estimators have analytical solutions that do not require the numerical searches associated with ℓ_1 regularization or Monte Carlo optimization common to many Bayesian estimators, as in Feng and He (2022), for example.

Other, more flexible machine learning methods may also be suitable for estimating the local ICs. The two key steps we have performed here are to construct economically motivated prediction features $D^j s$ and to cross-sectionally center the features in each group via fixed effects $\gamma^j D^j$. A key motivation for this approach rests on the assumption that the analyst's insight in choosing the groups is at least as helpful as generic machine learning methods applied to less carefully curated features. This obviously depends on the analyst's insight.

3.1 Generalized Ridge Regression

To estimate the coefficients, I extend the standard Ridge regression framework to allow both a full covariance penalty matrix and a non-zero shrinkage target $\boldsymbol{\beta}_0$. This generalization follows van Wieringen (2023), who show that the solution corresponds to the posterior mean in a Gaussian Bayesian model with prior mean and covariance.

For clarity, I will first apply this framework to the regression without time-varying fixed effects, where we want to apply regularization to all of the group-specific IC components. I will then extend the estimation methodology to include the time-varying fixed effects, which we do not want to regularize since they are meant to center the variables exactly.

In order to use mostly conventional regression symbols, let

- $\boldsymbol{y}$ denote the NT vector of excess returns r^{ex}_{it+1}
- $\boldsymbol{X}_\beta$ denote the $)NT \times k$ signal matrix with columns for group-level signal interactions
- $\boldsymbol{\beta}$ denote the k regression coefficients
- $\boldsymbol{\beta}_0$ denote the k-element shrinkage target
- $\widetilde{\boldsymbol{\Omega}}$ denote the (shrunk) $k \times k$ covariance matrix of the coefficient estimates
- $\lambda > 0$ be the overall Ridge penalty scalar

We set the shrinkage target

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \widehat{IC}, & 0, & \cdots, & 0 \end{bmatrix}'. \tag{8}$$

With this target, we shrink the intercept toward the overall IC and all group-level deviations toward zero. Absent a theoretical prior for the overall IC, we can run a first-stage panel regression without group effects to estimate $\widehat{IC}$.

As stated in van Wieringen (2023), the generalized Ridge regression objective is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}_\beta \boldsymbol{\beta}\|_2^2 + \lambda\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \widetilde{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}. \tag{9}$$

The generalizations are twofold. First, we adopt a non-zero shrinkage target. Second, we apply shrinkage based on the full inverse covariance of the estimates, $\widetilde{\boldsymbol{\Omega}}^{-1}$, unlike the uniform scalar shrinkage in regular Ridge regression.

Conveniently, this generalized problem still has a closed-form solution

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_\beta' \boldsymbol{X}_\beta + \lambda\, \widetilde{\boldsymbol{\Omega}}^{-1}\right)^{-1} \left(\boldsymbol{X}_\beta' \boldsymbol{y} + \lambda\, \widetilde{\boldsymbol{\Omega}}^{-1} \boldsymbol{\beta}_0\right). \tag{10}$$

This estimator shrinks noisy group effects toward zero while preserving a conventional, pooled IC estimate for the intercept. It balances data-driven estimation with prior beliefs about where the signal is effective. While the zero priors for the group effects are natural, we can certainly use any other values that we find reasonable for individual groups. If we believe the signal does not work for a particular signal, we could use $-\widehat{IC}$ as the prior for that group effect, so that the net IC is zero for the group.

This formulation penalizes directions in parameter space according to their estimated uncertainty: coefficients with high variance (i.e., low precision) receive stronger regularization. The use of $\widetilde{\boldsymbol{\Omega}}^{-1}$ allows the penalty to incorporate correlations between group-level coefficient estimates but the effect is moderated by the shrinkage in the covariance estimate.

Combined with hierarchical or overlapping group structures in $\boldsymbol{X}_\beta$, this method produces interpretable and stable coefficient estimates that respect both signal structure and estimation noise.

The estimate is equivalent to the Bayesian regression estimate in Chow (1983) with a normally distributed prior. The prior has a mean equal to $\boldsymbol{\beta}_0$ with variance proportional to $\widetilde{\boldsymbol{\Omega}}$. The latter is a natural empirical Bayes estimate of the prior variance.

The estimate is also equivalent to Bayesian forecast combinations or Bayesian model averaging. Black and Litterman (1992) apply this idea to asset allocation problems by combining investors' expected returns with market-implied expected returns. Hoeting, Madigan, Raftery, and Volinsky (1999) describe Bayesian model averaging for more general forecasts. In our application, each group overlaps with the entire investable universe and possibly with other subsets of the universe. The refined estimates in equation (10) average all applicable group-specific estimates with weights governed by the precision of the estimates. In most other applications, the estimates or predictions all cover the same universe. A main contribution here is that this logic can be applied to many separate groups by carefully considering the overlap among the groups.

In the absence of shrinkage, $\lambda = 0$, the estimates reduce to separately estimated, group-specific ICs. Such a framework appears in Eric H. Sorensen (2005), who estimate separate cross-sectional ICs across six overlapping groups of stocks. Such an approach seems reasonable when the number of groups is small relative to the number of observations but introduces a lot of estimation noise when there are many groups or some of the groups contain a small number of stocks.

While we can enforce positivity of the overall ICs during estimation, the associated inequality constraints prevent an analytical solution. The constrained quadratic program remains convex, however, and can be solved with standard numerical searches. Unless the postprocessing of the coefficients produces large coefficient changes, it is unlikely that postprocessing the coefficients is very different from imposing the positive IC constraints during estimation.[4]

3.2 Time-Varying Fixed Effects

We now extend the estimation framework to accommodate the time-varying group-level fixed effects while preserving the structured shrinkage applied to the group-dependent information coefficients (ICs). The idea is to estimate a single regression that includes a block of slope coefficients $\boldsymbol{\beta}$ for the group-dependent IC components, which are subject to regularization, and a block of fixed-effect coefficients $\boldsymbol{\gamma}$ for the time-varying group-level fixed effects, which are treated as unregularized intercept terms.

In this regression, let

- $\boldsymbol{y}$ be the stacked vector of cross-sectional return ranks, as before.
- $\boldsymbol{X}_\beta$ be the design matrix of signal interacted with group indicators, as before.
- $\boldsymbol{X}_\gamma$ be the $NT \times m$ design matrix of time-varying group dummies used as fixed effects (e.g., size and earnings), with separate columns for each group-time combination.
- $\boldsymbol{X} = [\boldsymbol{X}_\beta \ \boldsymbol{X}_\gamma]$ be the full design matrix with dimension $NT \times (k+m)$.
- $\boldsymbol{\beta}$ be the vector of slope coefficients, as before.
- $\boldsymbol{\gamma}$ be the m-element vector of unpenalized fixed effect coefficients.
- $\boldsymbol{b} = \begin{bmatrix}\boldsymbol{\beta}' & \boldsymbol{\gamma}'\end{bmatrix}'$ be the full coefficient vector with $k+m$ elements.

For the coefficient estimates, we use a shrinkage target

$$\boldsymbol{b}_0 = \begin{bmatrix}\boldsymbol{\beta}_0' & \mathbf{0}'\end{bmatrix}'. \tag{11}$$

The target $\boldsymbol{\beta}_0$ is defined in equation (8). The shrinkage target for the fixed effects is not material because we don't apply shrinkage to these coefficients. As a result, any value is acceptable here.

[4] While some regularized regression tools can enforce positivity of the parameters, this is not suitable here. The constraint that the IC should be positive everywhere permits negative elements of $\boldsymbol{\beta}$. We must permit negative elements in $\boldsymbol{\beta}$ so the regression can identify groups with below-average ICs.

We define the $(k+m) \times (k+m)$ regularization matrix $\boldsymbol{\Lambda}$ as a block-diagonal matrix

$$\boldsymbol{\Lambda} = \lambda \begin{bmatrix} \widetilde{\boldsymbol{\Omega}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{12}$$

where $\widetilde{\boldsymbol{\Omega}}$ is the (shrunk) covariance matrix for the signal-related coefficients. The fixed effects are unpenalized, as indicated by the zeros in the bottom-right block of $\boldsymbol{\Lambda}$.

The penalized least-squares objective for this regression is

$$\widehat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + (\boldsymbol{b} - \boldsymbol{b}_0)' \boldsymbol{\Lambda} (\boldsymbol{b} - \boldsymbol{b}_0) \right\}. \tag{13}$$

The solution has the same structure as before,

$$\widehat{\boldsymbol{b}} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{b}_0\right). \tag{14}$$

This formulation preserves the Bayesian shrinkage interpretation for the group-dependent signal effectiveness terms while incorporating exact centering of groups via time-varying fixed effects. By setting the regularization weights for the fixed effects to zero, we ensure that these terms are treated as conventional intercepts estimated without shrinkage.

Although we derived **??** analytically, like any regression estimator it involves a matrix inverse, which requires numerical solutions. As for all regression estimates, it is generally preferable to solve the normal equations instead of inverting a large matrix. **??** provides additional details for practical estimation of the panel regression.

Höchle, Schmid, and Zimmermann (2024) argue for the inclusion of firm-specific effects in panel regressions without group structure. It is mechanically straightforward to add firm-specific effects to the panel regressions here. This is especially true if we treat the firm-specific effects in line with the time-varying fixed effects and don't apply shrinkage. However, firm-specific fixed effects potentially introduce a large number of parameters and we should carefully consider whether they are important after we introduce group structure into the regressions. We will leave this for future work.

3.3 Sparse Signals or Strong Priors

For signals with limited coverage, we may choose to zero out the signal for an entire group. Under regularization, the corresponding group-specific coefficient estimate should be zero. In case this is not true, we should post-process the coefficients to enforce this condition.

Similarly, we may have a strong prior that a signal does not work at all for a particular group. We can zero out the signal for this group and, as a result, the regularized group-specific coefficient. If stocks in this group also appear in other groups, however, their overall IC will not be zero. This is statistically sensible since information from other, overlapping groups should shift our prior for the group in question.[5]

3.4 Testing Group Relevance

We can test whether a particular conditioning variable, firm size for example, is relevant for prediction. The test asks whether we can reject the null hypothesis that all slope coefficients associated with the dimension are jointly zero. This can be done via a standard F-test for the appropriate slope coefficients. If we can reject the null hypothesis, we can remove the corresponding dummies from the regression.

3.5 Shrinkage of the Coefficient Covariance Matrix

To improve stability in the generalized Ridge procedure, we can also apply shrinkage to the covariance matrix of the regression coefficient estimates. Specifically, the Ledoit and Wolf (2004) shrinkage approach shrinks the empirical covariance matrix $\widehat{\boldsymbol{\Omega}}$ toward a structured target matrix $\boldsymbol{T}$. I set the shrinkage target to the diagonal of the empirical covariance matrix, as discussed in Schäfer and Strimmer (2005),

$$\boldsymbol{T} = \operatorname{diag}(\widehat{\boldsymbol{\Omega}}). \tag{15}$$

The shrunk covariance matrix estimator is

$$\widetilde{\boldsymbol{\Omega}} = (1-\delta)\widehat{\boldsymbol{\Omega}} + \delta\boldsymbol{T}, \tag{16}$$

where $\delta \in [0,1]$ is the shrinkage intensity.

To set δ, we can use a simple rule that balances the number of coefficients k with the total sample size N, defined as

$$\delta = \min\left(1, \frac{k}{N}\right). \tag{17}$$

[5] If we think our prior is unmovable, we can post-process the estimates to enforce the zero IC. For overlapping groups, this involves an adjustment to all of the coefficients, similar to positivity constraints for the overall ICs.

This rule ensures more aggressive shrinkage when the number of coefficients is large relative to the number of observations. The resulting matrix $\widetilde{\Omega}$ is used in the generalized Ridge optimization step to weight deviations from the original coefficient estimates based on their statistical precision.

It seems sensible to use this shrunk estimate of the covariance matrix in the post-processing step that ensures positive ICs.

3.6 Choice of Baseline Groups

To help regularization discover sparsity, we can pre-process the data to find a "typical" group, which we then choose as the omitted category. If we omit an unusual group, it is likely that the regression will estimate IC deviations for all of the other groups, even though they may be very similar to each other. Omitting a typical group makes it more likely that the deviations in the other groups are immaterial.

Specifically, for each categorical group (e.g., sector), we compute groupwise ICs $\widehat{\beta}_g$ by regressing $r_{i,t+1}$ on $s_{i,t}$ within group g. Let n_g be the number of observations in group g, and let $N = \sum_g n_g$ be the total number of observations. The pooled average IC is

$$\widehat{\beta}_{\text{pooled}} = \frac{1}{N} \sum_g n_g \widehat{\beta}_g. \tag{18}$$

We choose the reference group g_0 to minimize $|\widehat{\beta}_g - \widehat{\beta}_{\text{pooled}}|$, so the omitted group is the most representative, increasing the likelihood that small deviations for other groups are shrunk to zero by regularization.[6]

3.7 Comparison with Traditional Factor Regressions

An alternative to our group-structured IC framework is to model the signal as an alpha factor within a generalized Fama and MacBeth (1973) cross-sectional factor regression framework. In this approach, signal portfolios and returns are allowed to vary across observable stock groups by including interactions between the signal and group dummies (e.g., signal × sector, signal × size group). Eric H. Sorensen (2005) follow this approach without applying shrinkage. Hellum, Pedersen, and Rønn-Nielsen (2024) apply this idea to a global factor model, where the groups are countries. When applying shrinkage to the estimates, they find that global components dominate the country-specific factors.

This approach faces significant limitations, To account for heterogeneous signal performance, one must include a large number of interaction terms

[6] Without regularization, the choice of the baseline groups has no effects on the model fit or the model predictions.

across combinations of sectors, industries, sizes, and other classifications. This leads to an explosion in the number of parameters, especially when many groups intersect. For example, if the factor regression contains 50 alpha factors and 100 groups, there are $50 \times 100 = 5,000$ group-specific factor returns.[7] Without regularization, this makes the model vulnerable to overfitting and instability. This is especially true if the regressions are run in pure cross-sections, period-by-period, as is common. Under this method, standard estimation may not be feasible. Moreover, the associated covariance matrix of factor returns is very challenging to estimate. In the case of 50 alpha factors across 100 groups, the 5,000 factor returns give rise to more than 12 million covariance parameters.

When we apply regularization to such cross-sectional regressions, they can produce inconsistent patterns of sparsity over time, especially under ℓ_1 regularization. When regularization sets a group-specific coefficient to zero in a given period, it is unclear how to treat the associated factor return: should it be recorded as zero, or as missing? Both choices introduce distortions, either biasing the return series or complicating downstream covariance estimation and signal combination.

In contrast, our proposed framework models cross-sectional signal effectiveness directly using a structured regression of ranked future returns on ranked signal values, where the slope varies additively by group. We apply regularization to a single, time-pooled model, yielding consistent and interpretable shrinkage across groups. This avoids the need to define and maintain thousands of group-level interaction terms and eliminates ambiguity about missing or zero factor returns.

Importantly, our model does not require explicit interaction between every signal and every group combination. Instead, deviations from a global baseline IC are regularized toward zero in a unified regression. This naturally accommodates sparse group-level structure without the combinatorial burden of modeling every possible cross-term.

The contextual adjustments to signals can use different contexts, or groups, for different signals. For example, we can allow the IC to vary across liquidity groups for one signal and across measures of analyst attention for another signal. These differences are isolated to each signal and do not complicate the factor model.

Thus, while the traditional factor regression approach may seem familiar, it lacks a coherent and scalable method for incorporating persistent

[7] Cong, Feng, He, and Li (2023) use Bayesian shrinkage to collapse some of these groups based on the estimated differences in the models for each group. This can reduce the number of separate groups when they are not empirically important.

group heterogeneity in signal effectiveness. Our structured IC regression addresses this problem directly, thereby preserving interpretability, enabling regularization, and ensuring consistency across time.

3.8 Historical Simulations

To avoid look-ahead bias in historical simulations, it is vital that we estimate this model on past data before making forecasts for future returns.

It would be a mistake to use the full history of the signal to estimate the parameters of the group-structured regression before running historical simulations. If we do this, we give additional weight to groups where the signal has done especially well. Of course, this will improve historical simulations. Unfortunately, it is unrealistic.

Fortunately, no such confusion can arise in making live return forecasts, where we are free to use all historical data in order to estimate the coefficients.

4 Portfolio Construction

From the contextual signal, we can derive contextual alphas, expected returns, and then use them in portfolio construction. Although this process is straightforward, it is helpful to discuss some details of this process and highlight some consequences of the contextual alphas.

4.1 Expected Returns from Estimated ICs

From the regression above, after enforcing positive ICs, the estimated slope for stock i is

$$\begin{aligned}\widetilde{\beta}_i = \widetilde{\beta}_0 &+ \sum_{g \neq g_0} \widetilde{\beta}_g^{\text{sec}} D_{ig}^{\text{sec}} + \sum_{h \neq h_{g(i)}} \widetilde{\beta}_h^{\text{ind}} D_{ih}^{\text{ind}} \\ &+ \sum_{d \neq d_0} \widetilde{\beta}_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \widetilde{\beta}_e^{\text{event}} D_{ie}^{\text{event}}.\end{aligned} \tag{19}$$

Then, the expected returns are

$$E_t r_{i,t+1} = \widetilde{\beta}_i s_{i,t}. \tag{20}$$

Although the ICs are estimated using ranks, expected returns for portfolio optimization are derived by scaling signal values by group-specific ICs, translating rank-based effectiveness into return forecasts.

The expected returns here intentionally omit the time-series average returns from the fixed effects. The role of the fixed effects is to identify the correct slope coefficients for the signal. If we believe that there are predictable average return differences across the groups, those can be captured in a separate signal.

4.2 Mapping Coefficients to Stock-Level ICs

To map each signal to the corresponding expected return, it is helpful to expand the group-wise ICs to security-wise ICs.

Let $\boldsymbol{\beta}$ be the k-element vector of estimated group-structured coefficients, and let $\boldsymbol{s}_t$ be the n-element vector of signal values for n stocks at time t.[8]

Define an $n \times k$ matrix $\boldsymbol{Z}$, where each row corresponds to a stock and each column corresponds to a coefficient in $\boldsymbol{\beta}$. The matrix $\boldsymbol{Z}$ encodes the group membership of each stock

$$Z_{ij} = \begin{cases} 1 & \text{if stock } i \text{ belongs to group } j \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Each row $\boldsymbol{Z}_i$ contains the dummy variables (including β_0) associated with stock i, and allows us to write the IC for stock i as

$$\rho_i = \boldsymbol{Z}_i \boldsymbol{\beta}. \tag{22}$$

This yields the stock-specific scale factors, or information coefficients,

$$\boldsymbol{\rho} = \boldsymbol{Z}\boldsymbol{\beta}. \tag{23}$$

Then, the expected returns are given by the elementwise product

$$\begin{aligned} E_t \boldsymbol{r}_{t+1} &= \boldsymbol{\rho} \odot \boldsymbol{s}_t \qquad (24) \\ &= (\boldsymbol{Z}\boldsymbol{\beta}) \odot \boldsymbol{s}_t. \qquad (25) \end{aligned}$$

This expression shows how the group-structured regression yields heterogeneous expected returns by applying different ICs to different stocks, based on their characteristics.

The map shows that the expected returns are not a rescaled version of the original signal. We apply different scale factors to different groups of stocks, depending on the group-wise efficacy of the signal. This can easily change the overall rank of the signal. For example, stock 1 can have a raw signal score $s_1 = 0.4$ and stock 2 can have a raw signal score $s_2 = 0.6$. If $\widetilde{\beta}_1 = 0.75$ and $\widetilde{\beta}_2 = 0.25$, the expected return for stock 1 is higher than for

[8] Although n may vary over time, we suppress t subscripts on the symbols affected by this: $n, \boldsymbol{Z}$, and $\boldsymbol{\rho}$.

stock 2, 0.30 versus 0.15, even though the raw signals implied the opposite order.

4.3 Optimal Portfolio: Single Signal

Let $\boldsymbol{\alpha} = \boldsymbol{\rho} \odot \boldsymbol{s}$ be the vector of expected returns from the signal, defined in equation (24), and $\boldsymbol{\Sigma}$ the residual covariance matrix from a factor model excluding the signal. Let $\boldsymbol{B}$ be the factor loading matrix for a collection of risk factors.

We solve the mean-variance portfolio optimization problem

$$\max_{\boldsymbol{w}} \quad \boldsymbol{w}'\boldsymbol{\alpha} - \frac{1}{2}\lambda \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \tag{26}$$

$$\text{s.t.} \quad \boldsymbol{w}'\boldsymbol{B} = 0. \tag{27}$$

The analytical solution for the optimal portfolio weights is

$$\widehat{\boldsymbol{w}} = \frac{1}{\lambda}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\alpha} \tag{28}$$

$$= \frac{1}{\lambda}\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \tag{29}$$

with

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}'. \tag{30}$$

This expression projects the unconstrained optimal portfolio onto the subspace orthogonal to the risk factors, thereby ensuring factor neutrality. The matrix $\boldsymbol{P}$ is the appropriate projection matrix. This yields a signal-exploiting portfolio that remains neutral to risk factors where signal strength varies across stock groups.

Portfolios based on the contextual alphas, $\boldsymbol{\alpha} = \boldsymbol{\rho} \odot \boldsymbol{s}$, deviate from portfolios based on the original signals, $\boldsymbol{s}$, in two ways. First, the contextual alphas have larger scale in groups where the alphas are more effective. The corresponding portfolios have larger gross exposure in those groups than the portfolios based on $\boldsymbol{s}$. Second, the contextual alphas can change the relative ranking of the stocks, as in the example above. Therefore, portfolios based on $\boldsymbol{\alpha}$ have different security rankings than portfolios based on $\boldsymbol{s}$. Because we require ICs to be positive everywhere, however, the $\boldsymbol{\alpha}$ portfolio is likely to have material exposure to the original signal $\boldsymbol{s}$, $\widehat{\boldsymbol{w}}'\boldsymbol{s} > 0$.

Portfolios based on contextual alphas also differ from portfolios that are constructed separately in each group. For example, we might construct separate portfolios in each sector. Portfolio construction with contextual alphas tunes alphas by groups but uses the entire investable universe for hedging purposes. This can materially reduce portfolio risk in the presence of pervasive risk factors in $\boldsymbol{\Sigma}$ or $\boldsymbol{B}$ that cannot be hedged well within subsets of the investable universe.

4.4 Optimal Portfolio: Multiple Signals

We can easily define the weighted average of several alphas as a composite alpha. This composite alpha can then be used in portfolio optimization, as for the single signal in the portfolio above.

The main challenge is choosing good weights across signals that may have material correlation with each other. Solutions to this problem can be improved by including the individual alphas in a multi-factor model in order to derive the pure alphas. The pure alpha exposures are orthogonal to each other, which can simplify the allocation choices.

5 Multi-Factor Models with Contextual Signals

Once contextualized signals are mapped to expected returns, as in equation (24), they can be treated like any other conventional signal or factor. Conventional signals are applied uniformly to the entire investable universe. The expected returns in equation (24) apply uniformly to all stocks, even though the underlying signals $\boldsymbol{s}_t$ do not.

To integrate multiple signals, we can run group-aware regressions for each signal, as described above. We can use the local ICs to form expected returns. These expected returns are suitable for inclusion in a fundamental factor model with both risk and alpha factors. For this factor model, we can run cross-sectional regressions of stock returns on risk and alpha factors to obtain pure factor returns. Finally, we can estimate expected returns and covariances for the factor returns.

The tuned, contextual alphas do not require additional special treatment prior to inclusion in a multi-factor model. However, there are some common pre-processing steps we might apply to raw signals. While some of these do not distort the contextual alphas, others do and we should be be cautious in their application to the contextual alphas.

We commonly standardize factor scores before including them in the factor regressions. Standardizing the overall expected return is a simple rescaling that does not affect the relative scale or ranking or the alphas. As a result, we can apply overall standardization without undoing any of the contextual adjustments.

In some cases, we standardize factor scores within groups, such as industries. This can partially undo the effects of the contextual signal adjustments and should be done with extreme care. In fact, it is probably best avoided.

Some factor regressions rank factor scores before standardizing them. Ranking, whether within groups or in the full estimation universe, at least partially removes the contextual scale adjustments and is best avoided. The main motivation for focusing on ranks is that they are highly robust to outliers. If we start with ranked signals $s_{i,t}$ there should be no urgent need to rank the expected returns. The original ranking adjusted outliers.

As for conventional factors, the pure factor portfolios corresponding to the tuned alpha signals have different weights and exposures than the raw factors. This is the result of the mutual orthogonalization performed by the regression. If the raw factor exposures are not highly correlated with the other factors, the changes may be minor.

As for conventional factors, we obtain a single factor return per period, even though the contextual signal was tuned to different groups of stocks. This stands in stark contrast to models that include separate factors for different groups of stocks in the overall factor model.

Unlike traditional dummy-interacted alpha factors, contextual signals remain parsimonious and interpretable, avoiding proliferation of factors and factor returns.

6 Conclusion

This paper develops a contextual modeling framework for signal effectiveness by allowing the information coefficient (IC) of a trading signal to vary systematically across groups of stocks defined by categorical characteristics such as sector, industry, size, and event time. By expressing the IC as a group-structured linear function, the model captures persistent heterogeneity in signal performance across the investable universe.

The core insight is to interpret the slope in a rank-based regression of future returns on the signal as local IC components that may differ by group. This structured regression permits granular decomposition of signal effectiveness and yields interpretable group-specific contributions to predictive power.

To ensure stable estimation, the model applies generalized Ridge regularization to the group-level IC components, guided by a shrinkage penalty matrix based on coefficient estimate precision. A post-estimation adjustment can enforce non-negative ICs across all groups to further stabilize the predictions.

The estimation framework accommodates group-specific fixed effects, which are not regularized and act to center returns and signals within group-period combinations. The resulting model is a generalized Bayesian Ridge regression with informative priors on the IC components and flat priors on the fixed effects.

Expected return forecasts constructed from the estimated ICs are readily usable in both single-signal and multi-signal portfolio construction. These forecasts reflect both the signal value and the context in which it is observed. The resulting portfolios amplify the signal where it is most predictive and mute it where it is less reliable, thereby leading to improved risk allocation and more consistent portfolio performance.

Overall, the approach offers a coherent and practically implementable method for refining trading signals for improved portfolio performance in the presence of cross-sectional heterogeneity. The approach combines statistical rigor with economic intuition and economically motivated constraints to derive interpretable signals that are flexible but more parsimonious than competing approaches.

7 References

Black, Fischer, and Robert Litterman, 1992, Global portfolio optimization, *Financial Analysts Journal* 48, 28–43.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, 2011, Robust inference with multiway clustering, *Journal of Business & Economic Statistics* 29, 238–249.

Campbell, John Y., and Samuel B. Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *Review of Financial Studies* 21, 1509–1531.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2024, Deep learning in asset pricing, *Management Science* 70, 714–750.

Chow, Gregory C., 1983, *Econometrics* (McGraw Hill, New York, NY).

Cong, Lin W., Guanhao Feng, Jingyu He, and Junye Li, 2023, Uncommon factors and asset heterogeneity in the cross section and time series, Working paper, Johnson College of Buiness, Cornell University, Ithaca, NY.

Eric H. Sorensen, Edward Qian, Ronald Hua, 2005, Contextual fundamentals, models, and active management, *Journal of Portfolio Management* 32, 23–36.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Feng, Guanhao, and Jingyu He, 2022, Factor investing: A Bayesian hierarchical approach, *Journal of Econometrics* 230, 183–200.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting characteristics nonparametrically, *Review of Financial Studies* 33, 2326–2377.

Gaure, Simen, 2013, Ols with multiple high–dimensional category variables, *Computational Statistics & Data Analysis* 66, 8–18.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

Hanauer, Matthias X., Amar Soebhag, Marc Stam, and Tobias Hoogteijling, 2025, Do machine learning models need to be sector experts?, Working paper, TUM School of Management, Munich, Germany.

Hellum, Oliver, Lasse Heje Pedersen, and Anders Rønn-Nielsen, 2024, How global is predictability? The power of transfer learning, Working paper, Copenhagen Business School, Copenhagen, Denmark.

Höchle, Daniel, Markus Schmid, and Heinz Zimmermann, 2024, Does unobservable heterogeneity matter for portfolio-based asset pricing tests?, Working paper, FHNW School of Business, Basel, Switzerland.

Hoerl, Arthur E., and Robert W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12, 55–67.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky, 1999, Bayesian model averaging: A tutorial, *Statistical Sciences* 14, 382–417.

Hou, Kewei, Chen Xue, and Lu Zhang, 2018, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Howard, Clint, 2024, Choices matter when training machine learning models for return prediction, *Financial Analysts Journal* 80, 81–107.

Jagannathan, Ravi, and Tongshu Ma, 2003, Risk reduction in large portfolios: Why imposing the wrong constraints helps, *Journal of Finance* 58, 1651–1683.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Ledoit, Olivier, and Michael Wolf, 2004, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88, 365–411.

Li, Bin, Alberto Rossi, Xuemin Yan, and Lingling Zheng, 2025, Machine learning from a "universe" of signals: The role of feature engineering, Working paper, Georgetown University, Washington, DC.

Müller, Karsten, and Simon N M Schmickler, 2025, Interacting anomalies, *The Review of Asset Pricing Studies* 15, 162–216.

Schäfer, Juliane, and Korbinian Strimmer, 2005, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology* 4, Article 32.

Shen, Zhouyu, and Dacheng Xiu, 2024, Can machines learn weak signals?, Working paper, Booth School of Business, University of Chicago, Chicago, IL.

Thompson, Samuel B., 2011, Simple formulas for standard errors that cluster by both firm and time, *Journal of Financial Economics* 99, 1–10.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B* 58, 267–288.

Tibshirani, Ryan J., and Jonathan Taylor, 2011, The solution of the generalized Lasso, *Annals of Statistics* 39, 1355–1371.

van Wieringen, Wessel, 2023, Lecture notes on Ridge regression, Working paper, University of Amsterdam, Amsterdam, The Netherlands.

Zou, Hui, and Trevor Hastie, 2005, Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320.

A Estimation Details

This appendix discusses two potential estimation issues and their solutions. First, although the generalized Ridge estimator in the main text is a conventional regression estimator, it may involve the inverse of a very large matrix and can become computationally taxing. Second, the covariance matrix of the slope estimates may not be clearly defined if some of the groups are always empty.

A.1 Coefficient Estimation

The generalized Ridge regression for a panel with m groups and T periods includes mT time-varying fixed effects. For example, for 100 groups and 1,000 periods, there are 100,000 fixed effects. These coefficients are nuisance parameters in the sense that we generally don't care about their values or standard errors. But their presence in the design matrix can make direct numerical computation of equation (14) inefficient or even infeasible.

One approach is to solve the associated normal equations:

$$\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda}\right)\widehat{\boldsymbol{b}} = \boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{b}_0. \tag{31}$$

This system is sparse and can be handled efficiently using sparse matrix libraries.

An alternative and often more efficient approach is to treat the time-varying fixed effects as nuisance parameters and remove their influence via demeaning. That is, we subtract the group-wise means from both the dependent and independent variables in each period, which guarantees that the corresponding fixed effects are zero.

If the groups are disjoint, such demeaning can be done group-by-group in a straightforward manner. However, if the groups overlap, sequential demeaning is invalid because adjusting for a later group can undo the mean-centering of earlier groups. Instead, we must simultaneously solve for the group means in each period.

Let $\boldsymbol{D} \equiv \boldsymbol{X}_\gamma$ denote the group dummy matrix, with n rows (one per observation) and m columns (one per group-time fixed effect).[9] The matrix $\boldsymbol{D}$ is very sparse, with a small number of nonzero entries per row.

For each column variable $\boldsymbol{z} \in \{\boldsymbol{y}, \boldsymbol{X}_{\beta,1}, \ldots, \boldsymbol{X}_{\beta,k}\}$, we compute the m-dimensional group mean vector γ_z by solving the sparse linear system

[9] The matrix $\boldsymbol{D}$ is not related to the vectors $\boldsymbol{d}_j$ we defined in the main text. They represent entirely different dummy variables.

$$\boldsymbol{D}'\boldsymbol{D}\,\gamma_z = \boldsymbol{D}'\boldsymbol{z}. \tag{32}$$

The demeaned column is then

$$\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{D}\gamma_z. \tag{33}$$

Because the matrix $\boldsymbol{D}'\boldsymbol{D}$ is the same for all right-hand sides, we can factor it once using sparse Cholesky or $LDL^\top$ decomposition and reuse the factorization for all variables. Alternatively, we can solve equation (32) using an iterative sparse least-squares solver.[10]

This approach yields exact within-group residuals even when the groups overlap, and is computationally efficient for large-scale problems due to the sparsity of the system.

Gaure (2013) recommends alternating projections to solve systems like equation (32). This method is extremely fast when group overlaps are limited. For heavily overlapping groups, direct solution of the sparse linear system is often faster.

Compared to solving the full regression problem, the speedup comes from two sources. First, demeaning amounts to solving a smaller system of equations. The regression solves a $(k+m)\times(k+m)$ system of equations. Here, we solve two smaller systems: $k\times k$ and $m\times m$. The dimensional reduction of $2km$ can be very meaningful. Second, the system of equations for the dummy equations is highly sparse and can be solved more efficiently than a dense system of similar size.

A.2 Covariance Estimation

Since the covariance matrix of the slope coefficients is used in shrinkage and portfolio construction, we must ensure it is properly estimated. A primary concern is residual correlation, either over time or within groups.

We estimate the covariance matrix of the slope coefficients by first running a conventional linear regression without shrinkage, possibly after demeaning the data to account for time-varying fixed effects. The residuals from this regression are then used to compute the usual OLS covariance estimate

$$\boldsymbol{\Omega} = \sigma_\varepsilon^2(\boldsymbol{X}_\beta'\boldsymbol{X}_\beta)^{-1}, \tag{34}$$

where σ_ε^2 is the residual variance and $\boldsymbol{X}_\beta$ is the matrix of right-hand-side variables corresponding to the shrinkage-eligible slope coefficients. As the

[10] In Python, routines such as cg or minres in scipy.sparse.linalg implement these methods.

main text explains, it can be useful to apply shrinkage to this covariance before using it in the generalized Ridge regression.

Because group-wise means are removed each period, serial correlation in group effects is eliminated. For residuals of liquid asset returns, serial correlation is typically minimal. If it exists, it is likely absorbed by the signal and thus not present in the residuals. However, correlation of residuals within groups may still occur. To account for such structure, one can estimate the covariance matrix using multi-way clustered standard errors as proposed by Thompson (2011) and Cameron, Gelbach, and Miller (2011).

Another issue arises when some groups are always empty: No stocks ever appear in them due to data availability or screening. In such cases, one can handle the problem automatically during estimation. Columns of the design matrix corresponding to empty groups will consist entirely of zeros or missing values. These columns can be masked during estimation. For the affected slope coefficients, it is reasonable to assign a slope estimated of 0, a variance equal to a large number, and covariances equal to 0. This is consistent with Bayesian shrinkage and does not affect estimation elsewhere.

Finally, when using demeaning rather than explicit fixed effects, we must adjust the degrees of freedom in the covariance estimation. A full regression recognizes that m fixed effects are estimated and uses $NT - m - k$ degrees of freedom. If we demean and then regress, a naive estimate might assume $NT - k$ degrees of freedom. To correct for this, we scale the covariance matrix by $(NT - k)(NT - m - k)$. This adjustment ensures consistency with full fixed-effect estimation.