Contextual Alpha: Emphasizing Forecasts Where They Work Best

Ludger Hentschel

November 28, 2025

Abstract

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

We estimate the group-level IC components via a panel regression with generalized Ridge (empirical Bayes) shrinkage. This shrinkage pulls the global intercept toward the pooled IC, penalizes imprecise local effects in proportion to their estimated variance, and leaves time-varying fixed effects unpenalized to ensure proper centering of each group-period combination.

After estimation, we solve a covariance-weighted quadratic program that adjusts the coefficients 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 these stock-specific ICs by the raw signal values yields group-aware, contextual alphas that amplify the signal where it is effective and mute it where it is weak. These alphas can be used directly in portfolio optimization or added to a multi-factor risk model to produce orthogonalized factor exposures.

The procedure generalizes to multiple signals, each with its own contextual hierarchy, and accommodates overlapping group definitions without proliferating dummy-interaction factors.

www.ludgerhentschel.com ludger@ludgerhentschel.com

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 event time.

We 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 return forecasting and portfolio construction.

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.

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 multi-factor 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-sector rank of future returns for stock i at time $t+1$, and let $s_{i,t}$ denote the within-sector 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 sector. This standardization is appropriate for sector-neutral portfolios and for within-sector signal comparisons.

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, we 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.

3 Group-Structured Regressions for ICs

This section presents a regression framework for measuring signal effectiveness across groups. We assume sector-neutral portfolios throughout and z-score both signals and returns within sectors. This ensures comparability across stocks and avoids spurious group-level effects.

We use panel regressions to estimate information coefficients (ICs) that may vary across categorical groups such as industry, size, and event time. Our focus is on how signal effectiveness varies, not on whether groups have systematically different average returns. To that end, we carefully separate signal slopes (ICs) from group-specific fixed effects.

3.1 Within-Sector ICs

Let $r_{i,t+1}$ denote the within-sector z-scored return of stock i at time $t + 1$, and let $s_{i,t}$ denote the within-sector z-scored value of the signal. The baseline regression is

$$r_{i,t+1} = \beta_0 s_{i,t} + \varepsilon_{i,t}, \tag{2}$$

where β_0 is the global IC, the rank correlation between the signal and future returns across the full universe, assuming sector neutrality.

To allow the IC to vary across industries, we extend this to

$$r_{i,t+1} = \left(\beta_0 + \sum_{h \neq h_0} \beta_h^{\text{sec}} D_{ih}^{\text{sec}}\right) s_{i,t} + \varepsilon_{i,t}, \tag{3}$$

where D_{ih}^{sec} is a dummy indicating membership in sector h, and h_0 is the reference sector. The coefficients β_h^{sec} are interpreted as deviations from the baseline IC.

Because both $r_{i,t+1}$ and $s_{i,t}$ are standardized within sector, the estimated slopes can be interpreted directly as within-sector ICs.

3.2 Group Adjustments Without Within-Group Standardization

Now suppose we want to allow the IC to vary across size or event-time groups, without standardizing the signals within those groups. In that case, the regression takes the form

$$r_{i,t+1} = \left(\beta_0 + \sum_{d \neq d_0} \beta_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \beta_e^{\text{event}} D_{ie}^{\text{event}}\right) s_{i,t} + \varepsilon_{i,t}, \tag{4}$$

where D_{id}^{size} and D_{ie}^{event} are dummies for size and event-time groups. Since $s_{i,t}$ is not re-standardized within size or event-time groups, these coefficients capture modifications to the global IC, not true within-group ICs.

To prevent the regression from picking up spurious effects due to group-level mean shifts, we include time-varying fixed effects for size and event-time groups

$$r_{i,t+1} = \cdots + \sum_t \sum_d \gamma_{d,t}^{\text{size}} D_{id,t}^{\text{size}} + \sum_t \sum_e \gamma_{e,t}^{\text{event}} D_{ie,t}^{\text{event}} + \varepsilon_{i,t}. \tag{5}$$

This ensures that any persistent differences in group average returns are absorbed by the γ terms, not by the IC estimates.

We do not include time-varying fixed effects for sectors or industries, since the data have been z-scored within sectors. If that assumption is relaxed, corresponding fixed effects should be included.

If the goal is to estimate true ICs within size or event-time groups, we must first z-score both $r_{i,t+1}$ and $s_{i,t}$ within those groups. This approach is only needed if we want to interpret the slopes as true ICs within each size or event-time group. This standardization has the effect that we only compare each stock to its peers in the same group. In most cases, we prefer the previous model with global standardization and time-varying fixed effects, which is easier to estimate and interpret in a sector-neutral portfolio context.

3.3 Combined Specification

A general model can include both sector-specific and group-specific IC adjustments

$$
\begin{aligned}
r_{i,t+1} = \Big(\beta_0 &+ \sum_{g \neq g_0} \beta_g^{\text{sector}} D_{ig}^{\text{sector}} \\
&+ \sum_{d \neq d_0} \beta_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \beta_e^{\text{event}} D_{ie}^{\text{event}} \Big) s_{i,t} \\
&+ \sum_t \sum_d \gamma_{d,t}^{\text{size}} D_{id,t}^{\text{size}} + \sum_t \sum_e \gamma_{e,t}^{\text{event}} D_{ie,t}^{\text{event}} + \varepsilon_{i,t}.
\end{aligned} \tag{6}
$$

Here, $s_{i,t}$ is the signal, z-scored within sectors, the β coefficients adjust the IC by group, and the γ coefficients are fixed effects that control for group mean shifts over time.

This formulation cleanly separates signal slope heterogeneity (which enters the portfolio) from group-level mean shifts (which do not). It yields interpretable, stable estimates of contextual signal effectiveness.

Each estimated coefficient β_g^j captures the marginal IC of the signal in group g, relative to a reference group. The global IC for group g is

$$
\widehat{\text{IC}}_g = \beta_0 + \beta_g^j. \tag{7}
$$

We show this specification to demonstrate that we can apply different scales to the signal, whether the signal is z-scored within the groups or not. While the exact interpretation of the effect differs, 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.

3.4 Time-Varying Fixed Effects

When signals and returns are not fully centered within a group, cross-sectional regressions may conflate true signal effectiveness with group-level shifts in means. To prevent this, we include time-varying fixed effects for size and event-time groups.

These fixed effects take the form

$$\sum_t \sum_d \gamma_{d,t}^{\text{size}} D_{id,t}^{\text{size}} \quad \text{and} \quad \sum_t \sum_e \gamma_{e,t}^{\text{event}} D_{ie,t}^{\text{event}}, \tag{8}$$

where $D_{id,t}^{\text{size}}$ is a dummy equal to one if stock i belongs to size group d at time t, and similarly for event-time group e.

These dummies absorb group-specific shifts in average returns, ensuring that the estimated ICs (the slope coefficients on the signal) reflect differences in predictive strength, not persistent group-level biases.

Because signals and returns are already z-scored within sectors, we omit time-varying fixed effects for sectors. If that centering is not applied, corresponding fixed effects should be included.

This separation of slope coefficients (for ICs) from group intercepts (for return levels) ensures that signal effectiveness is not confounded by unrelated return structure across groups.

There are two standard approaches to estimating the fixed effects. The first simply includes the appropriate dummy variables in the regression and estimates the fixed effects along with the slope coefficients. This can become cumbersome in large panels, where we may have thousands of period-specific fixed effects. The second approach subtracts the group-specific means from independent and dependent variables. After this de-meaning, a regression without the fixed effects produces the same slope coefficients as the regression with the fixed effects. In case we wish to estimate the fixed effects, we can reconstruct the coefficients and their standard errors after the estimation. In the presence of a large number of fixed effects, this can be computationally more efficient.

3.5 Enforcing Positivity

Campbell and Thompson (2008) show that enforcing sign constraints on return forecasts can improve predictive accuracy.[1] Here, we 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.

This procedure applies whether the group coefficients reflect within-group ICs (e.g., sectors with standardized signals) or deviations from a global IC (e.g., size or event-time adjustments), provided the design matrix encodes group memberships appropriately.

Let $\widehat{\boldsymbol{\beta}}$ denote the vector of estimated slope coefficients and $\widetilde{\boldsymbol{\Omega}}$ the covariance matrix of the estimated slope coefficients.[2] 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.

The quadratic program

$$\min_{\widetilde{\beta}} \quad (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})' \widetilde{\boldsymbol{\Omega}}^{-1} (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \tag{9}$$

subject to

$$\boldsymbol{d}_j' \widetilde{\boldsymbol{\beta}} \geq 0 \quad \text{for all group combinations } j, \tag{10}$$

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.

4 Estimation

The slope coefficients in equation (6) intentionally do not carry time subscripts. While the IC and its components may change over time, the main objective

[1] 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.

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

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 and the Zou (2006) adaptive Lasso regression strongly favors parsimony by setting some coefficients exactly to zero. The ℓ_2 regularization in the Hoerl and Kennard (1970) Ridge regression shrinks estimates toward zero but does not collapse them there. The combination of ℓ_1 and ℓ_2 regularization in the Zou and Hastie (2005) Elastic Net regression allows for elements of both. The horseshoe penalty of Carvalho, Polson, and Scott (2010) similarly suppresses weak deviations while leaving large effects mostly unchanged.

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 here. In this setting, we have already selected a signal; the main effort is to refine that signal by allowing for local variation. It is not obvious that a large number of small local refinements – and the associated estimation noise – produce better forecasts than a small number of larger, more confident adjustments. This is ultimately an empirical question and the answer may depend on the signal.

The main challenge in estimating the regularized regression in equation (6) 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. We 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.

4.1 Generalized Ridge Regression

To estimate the coefficients, we 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, we 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. We 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

$$\boldsymbol{\beta}_0 = \left[\widehat{IC}, \quad 0, \quad \cdots, \quad 0\right]'. \tag{11}$$

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

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}_\beta \boldsymbol{\beta} \right\|_2^2 + \lambda \, (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \widetilde{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}. \tag{12}$$

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

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_\beta' \boldsymbol{X}_\beta + \lambda \, \widetilde{\boldsymbol{\Omega}}^{-1} \right)^{-1} \left(\boldsymbol{X}_\beta' \boldsymbol{y} + \lambda \, \widetilde{\boldsymbol{\Omega}}^{-1} \boldsymbol{\beta}_0 \right). \tag{13}$$

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 group, we could use $-\widehat{IC}$ as the prior for that group effect, so that the net IC prior 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 (13) 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 Sorensen, Hua, and Qian (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.[3]

4.2 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.

[3] 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.

- $\boldsymbol{X}_\gamma$ 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

$$\boldsymbol{b}_0 = \begin{bmatrix}\boldsymbol{\beta}_0' & \boldsymbol{0}'\end{bmatrix}' . \tag{14}$$

The target $\boldsymbol{\beta}_0$ is defined in equation (11). 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.

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

$$\boldsymbol{\Lambda} = \lambda \begin{bmatrix}\widetilde{\boldsymbol{\Omega}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \tag{15}$$

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

$$\widehat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left\{ \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right\|_2^2 + (\boldsymbol{b}-\boldsymbol{b}_0)'\boldsymbol{\Lambda}(\boldsymbol{b}-\boldsymbol{b}_0) \right\} . \tag{16}$$

The solution has the same structure as before,

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda})^{-1} \left(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{b}_0\right) . \tag{17}$$

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 equation (17) 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. Appendix B 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 leave this for future work.

4.3 Sparse Signals or Strong Priors

For signals with limited coverage, we may choose to zero out the signal for an entire group. With 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.[4]

4.4 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.

4.5 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 $\boldsymbol{T}$. We set the shrinkage target to the diagonal of the empirical covariance matrix, $\boldsymbol{T} = \text{diag}(\widehat{\boldsymbol{\Omega}})$, as discussed in Schäfer and Strimmer (2005).

[4] 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.

The shrunk covariance matrix estimator is

$$\widetilde{\boldsymbol{\Omega}} = (1-\delta)\widehat{\boldsymbol{\Omega}} + \delta\,\mathrm{diag}(\widehat{\boldsymbol{\Omega}}), \tag{18}$$

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

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

$$\delta = \min\left(\frac{k^2}{NT-k-m}, \underline{\delta}\right). \tag{19}$$

Here, $NT-m-k$ is the number of degrees of freedom: the number of stocks N times the number of periods in the panel T minus the number of coefficients we estimate, including the number of fixed effects m. This rule ensures more aggressive shrinkage when the number of correlation coefficients is large relative to the number of observations. The resulting matrix $\widetilde{\boldsymbol{\Omega}}$ is used in the generalized Ridge optimization step to weight deviations from the original coefficient estimates based on their statistical precision.

Especially for very large panels, the default shrinkage can become quite small and a minimum shrinkage, like $\underline{\delta} = 0.33$ can be advisable.

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

4.6 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 estimates 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 group-wise 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

$$\widehat{\beta}_{\text{pooled}} = \frac{1}{N}\sum_g n_g \widehat{\beta}_g. \tag{20}$$

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.[5]

4.7 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 $\times$ sector, signal $\times$ size group). Sorensen, Hua, and Qian (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 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.[6] 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

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

[6] 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.

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

4.8 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 improves the results in 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.

5 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.

5.1 Expected Returns from Estimated ICs

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

$$\begin{aligned}\widetilde{\beta}_i = \widetilde{\beta}_0 &+ \sum_{g \neq g_0} \widetilde{\beta}_g^{\text{sec}} D_{ig}^{\text{sec}} \\ &+ \sum_{d \neq d_0} \widetilde{\beta}_d^{\text{size}} D_{id}^{\text{size}} + \sum_{e \neq e_0} \widetilde{\beta}_e^{\text{event}} D_{ie}^{\text{event}}.\end{aligned} \tag{21}$$

Then, the expected returns are

$$E_t r_{i,t+1} = \widetilde{\beta}_i s_{i,t}. \tag{22}$$

Note that, by default, the signal $s_{i,t}$ is not standardized within size or event-time groups. For these groups, the fixed effects absorb average return differences, not signal normalization.

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.

5.2 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.[7]

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

$$Z_{ij} = \begin{cases} 1 & \text{if stock } i \text{ belongs to group } j \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

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

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

$$\rho_i = \boldsymbol{Z}_i \boldsymbol{\beta}. \tag{24}$$

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

$$\boldsymbol{\rho} = \boldsymbol{Z}\boldsymbol{\beta}. \tag{25}$$

Then, the expected returns are given by the elementwise product

$$E_t \boldsymbol{r}_{t+1} = \boldsymbol{\rho} \odot \boldsymbol{s}_t \tag{26}$$
$$= (\boldsymbol{Z}\boldsymbol{\beta}) \odot \boldsymbol{s}_t. \tag{27}$$

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

5.3 Optimal Portfolio: Single Signal

Let $\boldsymbol{\alpha} = \boldsymbol{\rho} \odot \boldsymbol{s}$ be the vector of expected returns from the signal, defined in equation (26), 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

$$\max_{\boldsymbol{w}} \quad \boldsymbol{w}'\boldsymbol{\alpha} - \frac{1}{2}\lambda \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \tag{28}$$
$$\text{s.t.} \quad \boldsymbol{w}'\boldsymbol{B} = 0. \tag{29}$$

The analytical solution for the optimal portfolio weights is

$$\widehat{\boldsymbol{w}} = \frac{1}{\lambda}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\alpha} \tag{30}$$
$$= \frac{1}{\lambda}\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \tag{31}$$

with

$$P = I - \Sigma^{-1} B \left(B' \Sigma^{-1} B \right)^{-1} B'. \tag{32}$$

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.

5.4 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. We can apply the standard weighting approach to the contextual alphas we estimate here.

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.

6 Multi-Factor Models with Contextual Signals

Once contextualized signals are mapped to expected returns, as in equation (26), they can be treated like any other conventional signal or factor. The expected return model in equation (26) covers the full investable universe, but the resulting expected returns are heterogeneous across stocks, reflecting group-specific differences in signal efficacy.

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 the contextual expected returns may remove important heterogeneity in IC strength across groups – the very feature these signals are designed to exploit. 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.

7 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, or event time. By expressing the IC as a group-structured linear function, the model captures persistent cross-sectional heterogeneity in signal performance.

The core insight is to interpret the slope in a rank-based regression of future returns on the signal as a local IC component 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 ICs, guided by a shrinkage penalty matrix that reflects the precision of the coefficient estimates. A post-estimation adjustment can enforce non-negative ICs across all groups, incorporating economically motivated constraints to stabilize predictions and avoid spurious bets.

The estimation framework accommodates time-varying group fixed effects, which are not regularized and serve to center returns and signals within group-period combinations. This produces a generalized Bayesian Ridge regression: informative priors are placed on the IC components, and flat priors on the fixed effects.

Expected return forecasts derived from the contextual ICs translate each signal into a stock-specific alpha. These contextual alphas reflect both the value of the signal and the context in which it appears. In portfolio construction, this leads to targeted signal amplification where the IC is high, and muted exposure where the IC is weak or near zero, thereby improving risk allocation and stabilizing performance across market conditions.

Overall, the approach offers a coherent and practically implementable method for refining trading signals in the presence of cross-sectional heterogeneity. It combines statistical rigor with economic structure and domain-informed constraints, delivering interpretable signals that are more flexible than traditional approaches but remain parsimonious and robust.

8 References

Black, Fischer, and Robert Litterman, 1992, Global portfolio optimization, *Financial Analysts Journal* 48 (5), 28–43.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, 2011, Robust inference with multiway clustering, *Journal of Business & Economic Statistics* 29 (2), 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 (4), 1509–1531.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott, 2010, The horseshoe estimator for sparse signals, *Biometrica* 97 (2), 465–480.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2024, Deep learning in asset pricing, *Management Science* 70 (2), 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.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47 (2), 427–465.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81 (3), 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 (1), 55–67.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky, 1999, Bayesian model averaging: A tutorial, *Statistical Sciences* 14 (4), 382–417.

Hou, Kewei, Chen Xue, and Lu Zhang, 2018, Replicating anomalies, *Review of Financial Studies* 33 (5), 2019–2133.

Howard, Clint, 2024, Choices matter when training machine learning models for return prediction, *Financial Analysts Journal* 80 (4), 81–107.

Jagannathan, Ravi, and Tongshu Ma, 2003, Risk reduction in large portfolios: Why imposing the wrong constraints helps, *Journal of Finance* 58 (4), 1651–1683.

James, Willard, and Charles Stein, 1961, Estimation with quadratic loss, in Jerzy Neyman, ed., *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 361–379.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48 (1), 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.

Morris, Carl N., 1983, Parametric empirical bayes inference: Theory and applications, *Journal of the American Statistical Association* 78 (381), 47–55.

Müller, Karsten, and Simon N M Schmickler, 2025, Interacting anomalies, *The Review of Asset Pricing Studies* 15 (2), 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 (1), Article 32.

Schwarz, Gideon, 1978, Estimating the dimension of a model, *Annals of Statistics* 6 (2), 461–464.

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

Sorensen, Eric H., Ronald Hua, and Edward Qian, 2005, Contextual fundamentals, models, and active management, *Journal of Portfolio Management* 32 (1), 23–36.

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

Tibshirani, Robert, 1996, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B* 58 (1), 267–288.

Tibshirani, Ryan J., and Jonathan Taylor, 2011, The solution of the generalized Lasso, *Annals of Statistics* 39 (3), 1355–1371.

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

Zou, Hui, 2006, The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101 (476), 1418–1429.

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 (2), 301–320.

A Portfolio Gains

In a slightly simplified setting, we can derive the Sharpe ratio gains from recognizing IC differences across groups, compared to enforcing a single IC for all stocks. We do this by comparing the Sharpe ratios for two investors with nearly the same information about future returns. The only difference is that one investor assumes that there is a single pervasive IC, while the other investor understands that ICs vary across groups of stocks.

We consider N stocks partitioned into G disjoint groups, each with an equal number of stocks. Within each group, signal values $\boldsymbol{s}_g$ are cross-sectionally standardized to have mean zero and unit variance. In the main text, we enforced this for sectors.

We assume that returns are driven by K risk factors with zero expected returns and a single alpha factor with positive expected returns. The group-neutral signal motivates net zero portfolios within each group. With enough assets in each group, the portfolios hedge out the K risk factors within each group. This leaves the alpha factor $\boldsymbol{s}$ as the single pervasive factor across the group portfolios.

In this setting, the portfolio optimization decomposes into a two-step process. The first step finds the optimal long-short portfolios within each group. The second step optimizes allocations across the G group portfolios.

A.1 Two Investors

The two investors see the same signals but use different assumptions about the ICs in each group

- Investor A assumes that there is a single IC, constant across all groups. We denote this vector by $\boldsymbol{\mu}_A = (IC, \dots, IC)'$.
- Investor B correctly assumes that ICs differ across groups, but are constant within each group. We denote this vector by $\boldsymbol{\mu}_B = (IC_1, \dots, IC_G)'$.

Both investors construct identical portfolios within each group, but their allocations across groups differ.

A.2 Portfolio Weights

Let $\boldsymbol{C}$ denote the covariance matrix of group portfolio returns. The optimal portfolio weights chosen by each investor are

$$\boldsymbol{w}_A \propto \boldsymbol{C}^{-1}\boldsymbol{\mu}_A, \qquad \boldsymbol{w}_B \propto \boldsymbol{C}^{-1}\boldsymbol{\mu}_B. \tag{33}$$

For convenience, we can normalize weights so that $\boldsymbol{w}'\boldsymbol{C}\boldsymbol{w} = 1$ (unit variance portfolios). This does not affect the portfolios' Sharpe ratio because the group portfolios are net zero portfolios and we assume free leverage.

A.3 Sharpe Ratios

Under the true distribution of returns, the average returns are $\boldsymbol{\mu}_B$, and the Sharpe ratio for each investor is

$$SR_B = \frac{\boldsymbol{w}_B'\boldsymbol{\mu}_B}{\sqrt{\boldsymbol{w}_B'\boldsymbol{C}\boldsymbol{w}_B}} = \sqrt{\boldsymbol{\mu}_B'\boldsymbol{C}^{-1}\boldsymbol{\mu}_B}, \tag{34}$$

$$SR_A = \frac{\boldsymbol{w}_A'\boldsymbol{\mu}_B}{\sqrt{\boldsymbol{w}_A'\boldsymbol{C}\boldsymbol{w}_A}} = \frac{\boldsymbol{\mu}_B'\boldsymbol{C}^{-1}\boldsymbol{\mu}_A}{\sqrt{\boldsymbol{\mu}_A'\boldsymbol{C}^{-1}\boldsymbol{\mu}_A}}. \tag{35}$$

By construction, investor B's Sharpe ratio is always greater than or equal to investor A's because investor A uses the wrong expected returns. The Sharpe ratios are equal only when $IC_g = IC$ for all groups.

We can define the gain from recognizing the different ICs as the ratio of the Sharpe ratios

$$\mathcal{G} \equiv \frac{SR_B}{SR_A} \tag{36}$$

$$= \frac{\sqrt{\boldsymbol{\mu}_B'\boldsymbol{C}^{-1}\boldsymbol{\mu}_B}}{\boldsymbol{\mu}_B'\boldsymbol{C}^{-1}\boldsymbol{\mu}_A}\sqrt{\boldsymbol{\mu}_A'\boldsymbol{C}^{-1}\boldsymbol{\mu}_A}. \tag{37}$$

A.4 Special Cases

The single-factor structure of the covariance and the equal size of groups allows us to write the covariance matrix as

$$\boldsymbol{C} = \sigma^2\left[(1-\rho)\boldsymbol{I}_G + \rho\boldsymbol{\iota}\boldsymbol{\iota}'\right], \tag{38}$$

where $\boldsymbol{I}_G$ is the identity matrix, $\boldsymbol{\iota}$ is a vector of ones, and ρ is the common correlation among the group portfolio returns.

The analytical inverse of $\boldsymbol{C}$ is

$$\boldsymbol{C}^{-1} = \frac{1}{\sigma^2(1-\rho)}\boldsymbol{I}_G - \frac{\rho}{\sigma^2(1-\rho)(1-\rho+\rho G)}\boldsymbol{\iota}\boldsymbol{\iota}'. \tag{39}$$

Using this form, the Sharpe ratios can be expressed as:

$$SR_B^2 = \frac{1}{\sigma^2(1-\rho)}\left(\sum_{g=1}^{G} IC_g^2 - \frac{\rho}{1-\rho+\rho G}\left(\sum_{g=1}^{G} IC_g\right)^2\right), \tag{40}$$

$$SR_A^2 = \frac{\left(\sum_{g=1}^{G} IC_g\right)^2}{\sigma^2 G(1-\rho+\rho G)}. \tag{41}$$

The ratio of the two Sharpe ratios simplifies to

$$\frac{SR_B^2}{SR_A^2} = \frac{G}{1-\rho}\left(\frac{S_2}{S_1^2}(1-\rho+\rho G)-\rho\right) \tag{42}$$

$$= \frac{1+CV^2}{1-\rho}(1-\rho+\rho G) - \frac{\rho G}{1-\rho}, \tag{43}$$

where $S_1 = \sum_g IC_g$, $S_2 = \sum_g IC_g^2$, and CV is the coefficient of variation for the group ICs.

In this form, we can see that

- If all ICs are equal, then $S_2/S_1^2 = 1/G$ and the ratio equals 1 (no gain).
- If ICs are heterogeneous, $S_2/S_1^2 > 1/G$, and investor B strictly outperforms investor A.
- Higher variation in ICs relative to the mean IC, a higher coefficient of variation, CV, increases the gains from recognizing heterogeneity.
- If $\rho = 0$, the squared gain is equal to $1 + CV^2$.
- As ρ increases, the effective number of independent groups $N_{\text{eff}} = G/\big(1+\rho(G-1)\big)$ shrinks, so the Sharpe ratios decline for both investors.[8]
- The relative Sharpe ratio SR_B/SR_A rises with ρ: higher correlations make it more costly to ignore group heterogeneity, so the measured gains from recognizing heterogeneity are larger.

A.5 Numerical Illustrations

We can use these results to compute numerical illustrations of the gains. We first show the unconstrained gains we derived above. In some cases, this produces short positions in the signal portfolios with the lowest ICs. These results are shown in table 1. The table confirms that for long-short portfolios only the spread in ICs matters, not their average. For highly correlated group portfolios, the gains become very large because investor B constructs long-short combinations of the group portfolios, which hedge the remaining factor risk from the signal very well. These very large gains are only available when Investor B is willing to short the group portfolios with the lowest ICs. This requires a great deal of conviction in the IC estimates and is unlikely to be realistic.

For comparison, we also show the gains that are available if we don't allow short positions in any of the group portfolios. We compute the associated optimal portfolios and gains numerically. These results are shown in table 2. Obviously, the available gains are lower than for the long-short portfolios

[8] The effective number of groups is the number of groups that produces the same equally-weighted variance as G in the case of $\rho = 0$.

Table 1: Gains from Group-Specific ICs in Unconstrained Portfolios

	Mean IC = 0.01				Mean IC = 0.05			
CV	G=5	G=10	G=50	G=100	G=5	G=10	G=50	G=100
Panel A: ρ = 0.0								
0.25	1.02	1.03	1.03	1.03	1.02	1.03	1.03	1.03
0.50	1.09	1.10	1.12	1.12	1.09	1.10	1.12	1.12
1.00	1.33	1.37	1.41	1.41	1.33	1.37	1.41	1.41
1.50	1.65	1.73	1.79	1.79	1.65	1.73	1.79	1.79
Panel B: ρ = 0.1								
0.25	1.04	1.06	1.18	1.32	1.04	1.06	1.18	1.32
0.50	1.14	1.21	1.61	2.00	1.14	1.21	1.61	2.00
1.00	1.48	1.69	2.72	3.60	1.48	1.69	2.72	3.60
1.50	1.92	2.28	3.92	5.28	1.92	2.28	3.92	5.28
Panel C: ρ = 0.5								
0.25	1.14	1.27	2.03	2.69	1.14	1.27	2.03	2.69
0.50	1.47	1.85	3.67	5.09	1.47	1.85	3.67	5.09
1.00	2.37	3.27	7.12	10.04	2.37	3.27	7.12	10.04
1.50	3.37	4.77	10.62	15.01	3.37	4.77	10.62	15.01
Panel D: ρ = 0.9								
0.25	1.79	2.45	5.34	7.52	1.79	2.45	5.34	7.52
0.50	3.13	4.58	10.53	14.95	3.13	4.58	10.53	14.95
1.00	6.02	9.00	20.99	29.84	6.02	9.00	20.99	29.84
1.50	8.97	13.46	31.47	44.75	8.97	13.46	31.47	44.75

The table reports Sharpe ratio gains from recognizing group-specific information coefficients (ICs) relative to using a single average IC across all groups.

For the unconstrained (long-short) case, the gains are computed directly from the analytical expressions

$$\mathcal{G} = \frac{\sqrt{\mu_B' C^{-1} \mu_B}}{\mu_B' C^{-1} \mu_A} \sqrt{\mu_A' C^{-1} \mu_A},$$

where $\boldsymbol{\mu}$ collects the group-level expected returns implied by the group ICs and $\boldsymbol{C}$ is the correlation matrix of group portfolio returns.

Group-specific ICs are constructed to match the specified coefficient of variation (CV). Specifically, ICs are evenly spaced on the normal quantiles to ensure symmetry around the mean, with average ICs of 0.01 and 0.05 reported. This ensures comparability across the long-short and long-only results.

across groups. Also, for long-only portfolios, the average IC has an effect on the available gains, in addition to the spread in ICs.

Although one might expect the average IC to affect the long-only gains (since low average ICs might constrain some group weights to zero), in practice the Sharpe ratios are scale-invariant: multiplying all ICs by a positive constant does not change the optimization outcome under non-negativity. As a result, the gain ratio SR_B/SR_A depends only on the dispersion of ICs

Table 2: Gains from Group-Specific ICs in Long-Only Portfolios

	Mean IC = 0.01				Mean IC = 0.05			
CV	G=5	G=10	G=50	G=100	G=5	G=10	G=50	G=100
Panel A: ρ = 0.0								
0.25	1.02	1.03	1.03	1.03	1.02	1.03	1.03	1.03
0.50	1.09	1.10	1.11	1.12	1.09	1.10	1.11	1.12
1.00	1.32	1.36	1.38	1.38	1.32	1.36	1.38	1.38
1.50	1.60	1.65	1.70	1.70	1.60	1.10	1.70	1.70
Panel B: ρ = 0.1								
0.25	1.04	1.06	1.15	1.21	1.04	1.06	1.15	1.21
0.50	1.14	1.20	1.41	1.52	1.14	1.20	1.41	1.52
1.00	1.43	1.56	1.97	2.19	1.43	1.56	1.97	2.19
1.50	1.77	1.96	2.56	2.88	1.77	1.96	2.56	2.88
Panel C: ρ = 0.5								
0.25	1.12	1.18	1.34	1.41	1.12	1.18	1.34	1.41
0.50	1.33	1.46	1.54	1.89	1.33	1.46	1.77	1.89
1.00	1.80	2.06	2.64	2.88	1.80	2.06	2.64	2.88
1.50	2.28	2.67	3.52	3.87	2.28	2.67	3.52	3.87
Panel D: ρ = 0.9								
0.25	1.27	1.35	1.50	1.56	1.27	1.35	1.51	1.56
0.50	1.57	1.74	2.05	2.17	1.57	1.74	2.05	2.17
1.00	2.19	2.52	3.16	3.39	2.19	2.52	3.16	3.39
1.50	2.80	3.31	4.26	4.62	2.80	3.31	4.26	4.62

The table reports Sharpe ratio gains from recognizing group-specific information coefficients (ICs) relative to using a single average IC across all groups.

For the long-only case, the gains are computed by numerically solving the portfolio optimization with nonnegativity constraints on group weights. This projection reduces the flexibility to go short groups with low ICs, thereby compressing the attainable gains.

Group-specific ICs are constructed to match the specified coefficient of variation (CV). Specifically, ICs are evenly spaced on the normal quantiles to ensure symmetry around the mean, with average ICs of 0.01 and 0.05 reported. This ensures comparability across the long-short and long-only results.

(CV) and the correlation structure, not on the mean IC level.

A.6 Estimation Uncertainty

The results above assume that investor B knows the group-specific information coefficients. In practice, the investor estimates these ICs with material noise. Clearly, the presence of estimation noise reduces the effective dispersion across groups. As a result, the attainable Sharpe ratio gain from recognizing group-specific ICs may be materially smaller in practice than in the examples above.

The presence of this estimation noise motivates our shrinkage estimators of the group-specific ICs. Our logic parallels the classic result of James and

Stein (1961) that shrinkage estimators dominate sample means in multivariate normal problems, and the broader development of empirical Bayes methods for hierarchical models as surveyed in Morris (1983). While it is difficult to quantify the exact loss in Sharpe ratio gains without additional distributional assumptions, the qualitative implication is clear: estimation error attenuates the benefits of recognizing heterogeneity. The magnitude of the reduction increases with the number of groups and decreases with the useful sample size.

A.7 IC Gains

In our simplified setting with G equally sized groups, where each group's returns and signals have been standardized within the group the group-specific information coefficient (IC) ρ_g reflects the local correlation between signal and return, unconfounded by heterogeneity in scale across groups.

Suppose we wish to estimate the overall IC of a combined forecast signal. Under the assumption of homogeneity, we would assign equal weights to each group and compute

$$\mathrm{IC}_{hom} = \frac{1}{\sigma_r}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\rho_g = \frac{\sqrt{G}\,\overline{\rho}}{\sigma_r}, \tag{44}$$

where $\overline{\rho} = \frac{1}{G}\sum_g \rho_g$ is the average group IC, and σ_r is the standard deviation of the pooled return vector.

In contrast, under heterogeneity, we can construct an improved combined signal by weighting each group signal proportionally to its IC: $a_g \propto \rho_g$. This yields an overall IC of

$$\mathrm{IC}_{het} = \frac{1}{\sigma_r}\left(\sum_{g=1}^{G}\rho_g^2\right)^{1/2}. \tag{45}$$

The gain in overall IC from recognizing IC heterogeneity is the ratio

$$\begin{aligned}
\mathrm{Gain}_{IC} &= \frac{\mathrm{IC}_{het}}{\mathrm{IC}_{hom}} \\
&= \left(\frac{\frac{1}{G}\sum_g \rho_g^2}{\left(\frac{1}{G}\sum_g \rho_g\right)^2}\right)^{1/2} \\
&= \left(1 + \frac{\mathrm{Var}[\rho_g]}{\mathbb{E}[\rho_g]^2}\right)^{1/2}
\end{aligned}$$

$$= \left(1 + CV^2\right)^{1/2}, \tag{46}$$

where CV is the coefficient of variation of the group ICs. This is identical to the gain in Sharpe ratio under the assumption that the signal portfolio returns are uncorrelated across groups.

It is tempting to think that the gain in Sharpe ratio and the gain in overall IC should behave similarly. Indeed, when the signal portfolio returns across groups are uncorrelated, the gains in Sharpe ratio and in overall IC are identical. However, the relationship diverges when these signal returns are positively correlated.

In this case, the level of the Sharpe ratio falls with return correlations due to diminished diversification. But the gain in Sharpe ratio from recognizing IC heterogeneity becomes larger. When returns are positively correlated, the equal-weight portfolio becomes increasingly similar to a uniform bet, and the benefit of concentrating in high-IC groups grows in relative terms. That is, the equal-weighted portfolio suffers more than the optimal one as return correlation increases, so the relative Sharpe improvement rises.

For the IC, it may seem that return correlations are irrelevant, since the expression for IC_{het} depends only on the ρ_g. However, this overlooks an important point: when signal returns are correlated, the estimated signal-return covariances $s_g' r_g$ are themselves statistically dependent across groups. That is, the estimation errors in the ρ_g are correlated.

If we account for this dependence correctly when estimating a single overall IC, we obtain the same estimator as IC_{het}. Thus, even though the IC expression appears not to depend on signal return correlations, the efficient aggregation of ICs implicitly reflects this structure. IC_{het} remains the efficient (minimum-variance) estimate of overall predictive strength.

B 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.

B.1 Coefficient Estimation

The generalized Ridge regression for a panel with p groups and T periods includes pT 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 (17) inefficient or even infeasible.

One approach is to solve the associated normal equations

$$(\boldsymbol{X}'\boldsymbol{X} + \Lambda)\,\widehat{\boldsymbol{b}} = \boldsymbol{X}'\boldsymbol{y} + \Lambda \boldsymbol{b}_0. \tag{47}$$

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 p 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 p-dimensional group mean vector $\boldsymbol{\gamma}_z$ by solving the sparse linear system

$$\boldsymbol{D}'\boldsymbol{D}\,\gamma_z = \boldsymbol{D}'\boldsymbol{z}. \tag{48}$$

The demeaned column is then

$$\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{D}\boldsymbol{\gamma}_z. \tag{49}$$

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 (48) using an iterative sparse least-squares solver.[10]

[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.

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

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 (48). 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 $(k+p)\times(k+p)$ system of equations. Here, we solve two smaller systems: $k \times k$ and $p \times p$. The dimensional reduction of $2kp$ 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.

In historical simulations with expanding or rolling windows, we can further speed calculations by finding the full-sample fixed effects once and then selecting the appropriate values for each period. Since the fixed effects are period-specific, this does not introduce any look-ahead bias.

B.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

$$\boldsymbol{\Omega} = \sigma_\varepsilon^2(\boldsymbol{X}_\beta'\boldsymbol{X}_\beta)^{-1}, \tag{50}$$

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 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 estimate 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 p fixed effects are estimated and uses $NT - k - p$ 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 $(NT - k)/(NT - k - p)$. This adjustment ensures consistency with full fixed-effect estimation.

B.3 Weighted regressions

In the main text, equation (13) states the generalized Ridge / empirical Bayes estimate of contextual ICs in an unweighted setting. When we run regressions using raw or standardized returns rather than return ranks, residual risk typically varies by stock and period, with estimates σ_{it}. We can construct the diagonal weight matrix $\boldsymbol{W} = diag(1/\sigma_{it}^2)$ and obtain the weighted generalization

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_\beta' \boldsymbol{W} \boldsymbol{X}_\beta + \lambda\, \widetilde{\boldsymbol{\Omega}}^{-1}\right)^{-1} \left(\boldsymbol{X}_\beta' \boldsymbol{W} \boldsymbol{y} + \lambda\, \widetilde{\boldsymbol{\Omega}}^{-1} \boldsymbol{\beta}_0\right). \tag{51}$$

All likelihood-based quantities should be defined in the same weighted (whitened) space. Let $\boldsymbol{X} = [\boldsymbol{X}_\gamma\ \boldsymbol{X}_\beta]$ partition unpenalized columns $\boldsymbol{X}_\gamma$ (e.g., time fixed effects or a global IC) and penalized group columns $\boldsymbol{X}_\beta$. Residualize under $\boldsymbol{W}$,

$$\widetilde{\boldsymbol{y}} = \boldsymbol{M}_X^W \boldsymbol{y}, \qquad \widetilde{\boldsymbol{X}}_\beta = \boldsymbol{M}_X^W \boldsymbol{X}_\beta, \tag{52}$$

with

$$\boldsymbol{M}_X^W = \boldsymbol{I} - \boldsymbol{X}_\gamma (\boldsymbol{X}_\gamma' \boldsymbol{W} \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma' \boldsymbol{W}, \tag{53}$$

and apply the penalized regression to $(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_\beta)$ with the same $\boldsymbol{W}$. The prior mean $\boldsymbol{\beta}_0$ (baseline IC) should be computed using the same weights, typically by a weighted regression or on whitened data, and the prior covariance $\widetilde{\boldsymbol{\Omega}}$ estimated from historical samples using the same procedure.

If we use return ranks in the regressions, weighting observations based on return volatility is generally not useful.

B.4 Estimating shrinkage intensity

Values of λ near one or zero effectively target maximum returns, at least in sample. With a large sample and stable return structures, this should translate into attractive out-of-sample returns.

At these low values of λ, however, the resulting portfolios are fairly concentrated and have elevated risk. The shrinkage parameter λ therefore governs the trade-off between portfolios with high expected returns and high risk, and portfolios with lower return and lower risk. Because return and risk decline at different rates, intermediate values of λ typically maximize the Sharpe ratio.

To select λ for maximum Sharpe ratio rather than maximum expected return or in-sample correlation, we can pursue one of two alternatives.

First, we can choose λ via cross-validation. For this, we partition the sample into standard estimation and validation subsets, estimate $\boldsymbol{\beta}$ in each estimation subsample, and evaluate the resulting portfolio Sharpe ratio as a function of λ in the corresponding holdout sample. Repeating this procedure across subsamples yields an average out-of-sample Sharpe ratio conditional on λ. The λ associated with the maximum Sharpe ratio provides a conceptually clean and data-driven choice of shrinkage intensity. The main appeal of this approach is that it directly targets predictive performance and implicitly penalizes overfitting during estimation. The main cost is computational: it requires repeated estimation of the panel regression across many overlapping subsamples. Especially when the regressions roll through time, this can be expensive.

Second, we can instead define an auxiliary in-sample objective that trades off the relative gain in the Sharpe ratio attributable to local IC flexibility against the effective degrees of freedom implied by λ. This approach is computationally efficient and conceptually related to information criteria such as the BIC. It does, however, require calibration of the penalty constant c to yield shrinkage intensities that perform well out of sample. We now describe this approach in detail.

As shown in appendix A, the squared ex-ante Sharpe ratio is proportional to $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}(\lambda) = \widetilde{\boldsymbol{X}}_\beta\widehat{\boldsymbol{\beta}}(\lambda)$ are the stock-specific expected excess returns derived from the contextual alphas and $\boldsymbol{\Sigma}_r$ is the excess-return covariance matrix. We can write the ex-ante squared Sharpe ratio as

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\alpha} = \widehat{\boldsymbol{\beta}}'(\lambda)\,\widetilde{\boldsymbol{X}}_\beta'\boldsymbol{\Sigma}_r^{-1}\widetilde{\boldsymbol{X}}_\beta\,\widehat{\boldsymbol{\beta}}(\lambda). \tag{54}$$

Define

$$\boldsymbol{A} = \widetilde{\boldsymbol{X}}_\beta'\boldsymbol{\Sigma}_r^{-1}\widetilde{\boldsymbol{X}}_\beta, \tag{55}$$
$$\boldsymbol{M} = \widetilde{\boldsymbol{X}}_\beta'\widetilde{\boldsymbol{X}}_\beta, \tag{56}$$
$$\boldsymbol{Q}(\lambda) = \left(\boldsymbol{M} + \lambda\,\widetilde{\boldsymbol{\Omega}}^{-1}\right)^{-1}, \tag{57}$$
$$\boldsymbol{u}(\lambda) = \widetilde{\boldsymbol{X}}_\beta'\widetilde{\boldsymbol{y}} + \lambda\,\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\beta}_0, \tag{58}$$

so that

$$\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{Q}(\lambda)\boldsymbol{u}(\lambda). \tag{59}$$

Note that $\boldsymbol{M} = \widetilde{\boldsymbol{X}}_\beta'\widetilde{\boldsymbol{X}}_\beta$ and $\boldsymbol{A} = \widetilde{\boldsymbol{X}}_\beta'\boldsymbol{\Sigma}_r^{-1}\widetilde{\boldsymbol{X}}_\beta$ arise from distinct quadratic forms. The matrix $\boldsymbol{M}$ corresponds to the curvature of the regression likelihood and therefore enters the generalized Ridge estimator $\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{Q}(\lambda)\boldsymbol{u}(\lambda)$. In contrast, $\boldsymbol{A}$ corresponds to the mean-variance portfolio objective and enters only in the evaluation of the squared ex-ante Sharpe ratio. The two matrices are not interchangeable.

Let

$$f(\lambda) = \boldsymbol{u}(\lambda)'\boldsymbol{Q}(\lambda)\,\boldsymbol{A}\,\boldsymbol{Q}(\lambda)\,\boldsymbol{u}(\lambda), \tag{60}$$

and

$$f_\infty = \lim_{\lambda\to\infty} f(\lambda) \tag{61}$$
$$= \boldsymbol{\beta}_0'\,\widetilde{\boldsymbol{X}}_\beta'\boldsymbol{\Sigma}_r^{-1}\widetilde{\boldsymbol{X}}_\beta\,\boldsymbol{\beta}_0, \tag{62}$$

corresponding to the Sharpe ratio proxy implied by the global or prior-only specification with no local IC adjustment. We then define the relative Sharpe ratio gain as

$$h(\lambda) = \frac{f(\lambda)}{f_\infty} - 1. \tag{63}$$

The penalized Sharpe-gain objective becomes

$$J(\lambda) = \underbrace{\left[\frac{f(\lambda)}{f_\infty} - 1\right]}_{\text{relative Sharpe ratio gain from local ICs}} - \pi \underbrace{\frac{1}{p}\,\mathrm{Tr}\big(\boldsymbol{A}\,\boldsymbol{Q}(\lambda)\big)}_{\text{flexibility penalty}} \tag{64}$$

$$= h(\lambda) - \pi\, g(\lambda), \tag{65}$$

where p is the number of shrinkage-eligible group coefficients in $\boldsymbol{\beta}$ and $\mathrm{Tr}(\boldsymbol{AQ}(\lambda))/p$ is the normalized effective degrees of freedom. The tuning constant $\pi \geq 0$ controls the parsimony penalty, analogous to the $\log n$ term in the Bayesian Information Criterion of Schwarz (1978).

This ratio formulation penalizes the relative improvement in the Sharpe ratio rather than its absolute change. Because $f(\lambda)/f_\infty$ measures the squared Sharpe ratio of the locally adjusted model relative to its global or prior-only baseline, the resulting measure is scale-free and comparable across signals. When local ICs add little value, $f(\lambda)/f_\infty \approx 1$, and the penalty term dominates. When local ICs materially improve risk-adjusted returns, $f(\lambda)/f_\infty$ rises, offsetting the penalty. Maximizing $J(\lambda)$ therefore selects λ to balance proportional Sharpe improvements against model flexibility, analogous to a regularized performance ratio.

Because $J(\lambda)$ is a smooth, one-dimensional function, it can be optimized efficiently by grid search or by a standard univariate algorithm such as Brent's method applied to $\log\lambda$. Analytical derivatives are available

$$\frac{df(\lambda)}{d\lambda} = 2\,\boldsymbol{u}'\boldsymbol{QAQ}\,\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\beta}_0 - \boldsymbol{u}'\Big(\boldsymbol{Q}\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{QAQ} + \boldsymbol{QAQ}\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{Q}\Big)\boldsymbol{u}, \tag{66}$$

$$\frac{dg(\lambda)}{d\lambda} = -\frac{1}{p}\,\mathrm{Tr}\big(\boldsymbol{A}\,\boldsymbol{Q}(\lambda)\,\widetilde{\boldsymbol{\Omega}}^{-1}\,\boldsymbol{Q}(\lambda)\big), \tag{67}$$

but numerical maximization is sufficient in practice.

At $\pi = 0$, maximizing $J(\lambda)$ yields λ_0, which maximizes the in-sample relative gain in Sharpe ratio from local ICs.[11] Introducing a positive penalty, $\pi > 0$, introduces stable shrinkage and discourages spurious local variation. A convenient baseline penalty can be set by slope matching at λ_0, giving $\pi_0 = -f'(\lambda_0)/g'(\lambda_0)$. Setting $\pi = \gamma\,\pi_0$ with $\gamma \approx 1.5$ typically yields robust shrinkage levels that improve out-of-sample Sharpe ratios by reducing effective model complexity.

Empirically, this relative Sharpe-gain criterion produces results consistent with cross-validation but at far lower computational cost. It commonly selects

[11] Since the estimation does not explicitly maximize the Sharpe ratio, λ_0 may differ from 0.

$\lambda > 1$, often in the range $\lambda \approx 3$–30, corresponding to modest but effective shrinkage toward the prior. The resulting portfolios are less concentrated across groups, exhibit more stable local IC estimates, and achieve higher out-of-sample Sharpe ratios despite slightly lower in-sample gains.

If $\boldsymbol{\Sigma}_r = \boldsymbol{I}$, then $\boldsymbol{A}$ reduces to $\widetilde{\boldsymbol{X}}_\beta' \widetilde{\boldsymbol{X}}_\beta$ and all expressions above specialize accordingly. Because returns are demeaned within groups each period, this approximation to $\boldsymbol{\Sigma}_r$ is reasonable, especially when regressions use return ranks rather than raw returns.

In practice, both approaches, cross-validation and the penalized relative Sharpe-gain criterion, often lead to similar shrinkage intensities. Cross-validation provides a conceptually clean benchmark, while the penalized Sharpe-gain criterion offers a fast and reproducible alternative for large-scale or rolling-panel applications.