Negative Idiosyncratic Returns in Optimized Portfolios

Ludger Hentschel

September 12, 2025

Abstract
Empirically, portfolios constructed via mean-variance optimization often exhibit negative idiosyncratic return contributions: the realized returns attributable to residual (non-factor) exposures are negative, on average, over time. This is especially surprising under the standard fundamental factor model, where residual returns are assumed to be zero-mean, uncorrelated with factors, and uncorrelated across assets.

We show that persistent deviations from zero idiosyncratic returns are indications of omitted factors and describe three ways of correcting this issue. Even without knowing the true omitted factors, we know that their component aligned with the direction of trade contains alpha. First, we can add an appropriate factor with alpha to the alpha factor model. Second, to the extent trading costs drive the deviations, improved alpha scaling may help. Third, we can add a constraint or penalty for idiosyncratic risk that is not part of the alpha factors. All of these methods can be implemented in standard portfolio optimization software.

www.ludgerhentschel.com   ludger@ludgerhentschel.com

Contents

1 Introduction

Optimized portfolios often exhibit negative idiosyncratic (residual) return contributions. The realized returns attributable to residual (non-factor) exposures are negative, on average, over time. This is surprising under the standard assumptions of a fundamental factor model, which treats residual returns as independent across assets, mean-zero, and uncorrelated with factor returns. Under those assumptions, residual exposures should not contribute to expected performance in either direction, so their average contribution should be close to zero.

If a manager's alphas extend beyond the span of the factor model's factor space, residual returns can legitimately contribute to performance, positively or negatively, because they reflect additional sources of systematic return. However, the puzzle is sharpest when the optimizer's target lies entirely in the factor space. This occurs, for example, when starting from a composite of pure factor portfolios and using the corresponding implied alphas. In the frictionless baseline, the optimal portfolio then coincides with the factor composite and earns zero idiosyncratic return by construction.

In practice, trading costs and portfolio constraints force deviations from the factor space. When average residual contributions are persistently negative, this indicates a systematic alignment between the residual exposures we actually hold and future residual returns. This is equivalent to saying that there is an omitted priced factor in residual space. This note formalizes that interpretation, proposes two statistical tests to distinguish "bad luck" from a real omitted factor, and presents practical remedies.

Shifts in factor weights due to costs or constraints can also occur. These remain in factor space and therefore do not produce idiosyncratic returns. They may still be important to monitor, but they are not the focus here. For methods to constrain relative factor exposures during portfolio optimization, see Hentschel (2025b). We focus exclusively on diagnosing and addressing the idiosyncratic return issue.

Reducing the impact of negative idiosyncratic returns is not free, it is likely to increase transaction costs. The objective is to find a better balance by showing the optimizer more realistic performance costs, via expected returns and risks, in addition to the transaction costs.

2 Factor Model and Implied Alphas

Let $\boldsymbol{r}_{t+1}$ denote the $(n \times 1)$ vector of asset returns from t to $t{+}1$. We assume a K-factor fundamental model

$$\boldsymbol{r}_{t+1} = \boldsymbol{X}_t \boldsymbol{b}_{t+1} + \boldsymbol{\varepsilon}_{t+1}, \tag{1}$$

where $\boldsymbol{X}_t$ is the $(n \times K)$ matrix of factor exposures at t; $\boldsymbol{b}_{t+1}$ is the $(K \times 1)$ vector of factor returns over $[t, t{+}1]$; and $\boldsymbol{\varepsilon}_{t+1}$ is the $(n \times 1)$ vector of residual returns. The residuals $\boldsymbol{\varepsilon}_{t+1}$ are assumed to be mean-zero, independent across assets, and uncorrelated with $\boldsymbol{b}_{t+1}$.

2.1 Factor Risk Model

The corresponding return covariance is

$$\boldsymbol{\Sigma}_t = \boldsymbol{X}_t \boldsymbol{\Omega}_t \boldsymbol{X}_t' + \boldsymbol{D}_t, \tag{2}$$

where $\boldsymbol{\Omega}_t$ is the $(K \times K)$ factor covariance matrix and $\boldsymbol{D}_t$ is the diagonal $(n \times n)$ matrix of residual variances. All common co-movement is captured by the factors; $\boldsymbol{D}_t$ measures the risk unique to each asset.

2.2 Pure Factor Portfolios

From $\boldsymbol{X}_t$ we can construct the K pure factor portfolios

$$\boldsymbol{W}_t = \boldsymbol{X}_t \left(\boldsymbol{X}_t' \boldsymbol{X}_t\right)^{-1} \tag{3}$$

or

$$\boldsymbol{W}_t = \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{X}_t \left(\boldsymbol{X}_t' \boldsymbol{\Gamma}_t \boldsymbol{X}_t\right)^{-1} \tag{4}$$

The first factor portfolios are the result of ordinary least squares factor regressions. The second factor portfolios are the result of weighted least squares factor regressions with weights $\boldsymbol{\Gamma}_t$. Common weights come from idiosyncratic risk, $\boldsymbol{\Gamma}_t = \boldsymbol{D}_t^{-1}$, or proxies for this, like market capitalization. Column j of $\boldsymbol{W}_t$ is the $(n \times 1)$ weight vector with unit exposure to factor j and zero exposure to all other factors. Any portfolio in factor space is a linear combination of these columns.

2.3 Target Portfolio

Let the $(n \times 1)$ vector $\widehat{\boldsymbol{w}}_t$ represent the manager's desired factor composite, a linear combination of pure factor portfolios in the columns of $\boldsymbol{W}_t$.

Sharpe (1974) shows that the implied alphas reproducing $\widehat{\boldsymbol{w}}_t$ under mean-variance optimization are

$$\widehat{\boldsymbol{\alpha}}_t = \lambda\, \boldsymbol{\Sigma}_t\, \widehat{\boldsymbol{w}}_t, \tag{5}$$

where $\lambda > 0$ is the risk-aversion parameter.

Using these alphas in a Markowitz (1952) mean-variance optimization without frictions yields exactly the factor composite,

$$\boldsymbol{w}_t^{\star} = \lambda^{-1} \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\alpha}}_t = \widehat{\boldsymbol{w}}_t. \tag{6}$$

Thus, the optimized portfolio equals the target factor composite in the absence of costs or constraints.[1]

2.4 Residual Exposures

We define the projection matrix onto factor space as

$$\boldsymbol{P}_X = \boldsymbol{X}_t (\boldsymbol{X}_t' \boldsymbol{X}_t)^{-1} \boldsymbol{X}_t' \tag{7}$$

or

$$\boldsymbol{P}_X = \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{X}_t (\boldsymbol{X}_t' \boldsymbol{\Gamma}_t \boldsymbol{X}_t)^{-1} \boldsymbol{X}_t' \boldsymbol{\Gamma}_t^{1/2}. \tag{8}$$

The first form is appropriate for ordinary least squares (OLS); the second for weighted least squares (WLS). In the remainder, we will refer to $\boldsymbol{P}_X$ generically and the reader should choose the form that corresponds to the factor model regressions.

The residual, non-factor, portion of a portfolio $\boldsymbol{w}$ is

$$(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w}. \tag{9}$$

This computation is equivalent to finding the residuals from cross-sectionally regressing the portfolio weights $\boldsymbol{w}$ on the exposures for all of the factors $\boldsymbol{X}$, either using ordinary or weighted least squares.

For any factor portfolio $\boldsymbol{w}_t$ in $\boldsymbol{W}_t$, including the target factor composite $\widehat{\boldsymbol{w}}_t$, $(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w} = \boldsymbol{0}$ because

$$(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{X}_t \left(\boldsymbol{X}_t' \boldsymbol{X}_t\right)^{-1} = \boldsymbol{0}. \tag{10}$$

[1] In general, the optimized portfolio is a levered version of the target portfolio unless we choose the the same coefficient of risk aversion in the optimization and when computing implied alphas.

This is also true for the weighted portfolios. For factor portfolios, the idiosyncratic exposures and return contributions are exactly zero.

3 Residual Returns from Costs and Constraints

With trading frictions, the optimizer solves the Markowitz (1952) mean-variance problem

$$\boldsymbol{w}_t^\star = \arg\max_{\boldsymbol{w}\in\mathcal{W}} \left\{ \boldsymbol{w}'\widehat{\boldsymbol{\alpha}}_t - \frac{\lambda}{2}\boldsymbol{w}'\boldsymbol{\Sigma}_t\boldsymbol{w} - \tau(\boldsymbol{w}, \boldsymbol{w}_{t-1}) \right\}. \tag{11}$$

Here, $\mathcal{W}$ encodes portfolio constraints (e.g., leverage limits, position bounds, sector limits, turnover caps); and $\tau(\boldsymbol{w}, \boldsymbol{w}_{t-1})$ measures total transaction costs as a function of the trade $\boldsymbol{w} - \boldsymbol{w}_{t-1}$.

In general, $\boldsymbol{w}_t^\star \neq \widehat{\boldsymbol{w}}_t$, so the residual portfolio

$$\boldsymbol{w}_t^\perp \equiv (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w}_t^\star \tag{12}$$

is nonzero, which introduces idiosyncratic returns. The idiosyncratic return contribution for $[t, t+1]$ is

$$r_{t+1}^{id} = (\boldsymbol{w}_t^\perp)'\boldsymbol{\varepsilon}_{t+1}. \tag{13}$$

If $\boldsymbol{\varepsilon}_{t+1}$ is truly iid, mean-zero, and independent of $\boldsymbol{w}_t^\perp$, then

$$E[r_{t+1}^{id}] = 0. \tag{14}$$

Changes in factor weights alone cannot generate idiosyncratic returns, since they remain in factor space. Only deviations into residual space can do so.

4 Omitted Factor Interpretation

Persistent $r_{t+1}^{id} < 0$ implies a systematic alignment between $\boldsymbol{w}_t^\perp$ and future $\boldsymbol{\varepsilon}_{t+1}$. That indicates the presence of a priced residual factor $\boldsymbol{w}_t^\perp$, to which the portfolio has systematic and costly exposure.

We can also link this omitted factor interpretation to trades. Define the trade gap between the target and the incoming portfolio

$$\boldsymbol{g}_t \equiv \widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}. \tag{15}$$

Project this into residual space

$$g_t^{\perp} \equiv (\boldsymbol{I} - \boldsymbol{P}_X)g_t. \tag{16}$$

This is the part of the desired trade that lies outside factor space. Costs and constraints often prevent us from fully closing $g_t^{\perp}$, so the implemented residual exposure $\boldsymbol{w}_t^{\perp}$ tends to align with it. Moreover, the exposure is likely negative because the portfolio does not trade all the way to the target $\widehat{\boldsymbol{w}}_t$ in the presence of costs and constraints.

If

$$E[(g_t^{\perp})'\boldsymbol{\varepsilon}_{t+1}] > 0, \tag{17}$$

then $g_t^{\perp}$ behaves like a priced residual factor with positive expected return. The optimizer implicitly holds negative exposure to this factor. It does so freely because it values the exposure at zero ex ante. But the portfolio incurs a systematic drag ex post because it has negative exposure to a factor with positive average return.

The main problem is not that there is an omitted factor but that the optimizer assigns zero expected return to an omitted factor with positive returns and sees relatively small risk in these exposures because they diversify across a large number of securities.

5 Tests for an Omitted Factor

We recommend two tests for omitted factors, starting from the same factor model as used in the risk model and portfolio construction.

5.1 Test 1: Time-Series Mean of Idiosyncratic Returns

Compute the realized idiosyncratic return each period

$$r_{t+1}^{id} = \boldsymbol{w}_t'\boldsymbol{\varepsilon}_{t+1}. \tag{18}$$

Test

$$H_0 : E[r_{t+1}^{id}] = 0 \tag{19}$$

using the sample mean $\bar{r}^{id}$ and a Newey and West (1987) t-statistic with lags matched to the portfolio holding period. A significantly negative mean indicates a systematic residual drag. Winsorization may be applied to reduce the impact of outliers.

5.2 Test 2: Full Fama-MacBeth with Gap Exposure

At each t, run the cross-sectional Fama and MacBeth (1973) regression

$$r_{i,t+1} = \boldsymbol{X}_{i,t}\boldsymbol{b}_t + x_{i,t}^{gap}\beta_t^{gap} + e_{i,t+1}, \tag{20}$$

where $r_{i,t+1}$ is the return on asset i; $\boldsymbol{X}_{i,t}$ is the $(1 \times K)$ factor exposure vector; and $x_{i,t}^{gap}$ is the residual-space gap exposure

$$\boldsymbol{x}_t^{gap} \equiv (\boldsymbol{I} - \boldsymbol{P}_X)(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}). \tag{21}$$

Using the projected gap ensures factor returns and pure factor portfolios remain unchanged relative to a regression without it. We should run this augmented regression using the same weighting method we use in the main factor regressions and in the projection $\boldsymbol{P}_X$. We should also apply the same cross-sectional standardization to $\boldsymbol{x}_t^{gap}$ that we apply to the other factors.

Test

$$H_0 : E[\beta_t^{gap}] = 0 \quad \text{vs.} \quad H_A : E[\beta_t^{gap}] > 0 \tag{22}$$

with Newey and West (1987) standard errors on the $\{\widehat{\beta}_t^{gap}\}$ series.

Including $\boldsymbol{x}_t^{gap}$ alongside $\boldsymbol{X}_t$ controls for all model factors and isolates incremental pricing of the gap exposure. Projecting prevents leakage into factor returns and avoids multicollinearity.

A significantly positive average value of β_t^{gap} indicates that residual trade gaps, the part of the desired trade blocked by costs or constraints, are systematically aligned with future residual returns. This supports the interpretation of a positively priced omitted factor in residual space.

The time-series test measures whether the portfolio systematically gains or loses from residual exposures and should use the actual (unstandardized) portfolio weights. The Fama-MacBeth regression tests whether gap exposure is priced in the cross-section and may use standardized (z-scored) exposures for numerical stability and interpretability. The two tests answer related but distinct questions: the first focuses on realized performance, the second on cross-sectional pricing.

6 Remedies

There are occasions when noticing an omitted factor provides an impetus for modeling factors that we suspected previously. Identifying the omitted factor and including it in the fundamental factor model is the preferred solution to this problem.

Generally, however, the factors are omitted because we don't know what they are. Without identifying the factors, we cannot include them in the factor model. Here we focus on solutions to the problem of negative idiosyncratic returns that do not identify the omitted factor.

6.1 Auxiliary Gap Factor

Although we may not have a fundamental interpretation of the omitted factor, we know that the difference between the target portfolio and the actual portfolio has historically aligned with the omitted factor.

We can use the projected trade gap in residual space

$$\boldsymbol{x}_t^{gap} = (\boldsymbol{I} - \boldsymbol{P}_X)(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}) \tag{23}$$

as a proxy for the omitted factor.

It is common to standardize factor exposures to have unit cross-sectional variance (or norm) at each time step. When working with the projected trade gap $\boldsymbol{x}_t^{gap}$ as a proxy factor, we recommend standardizing it as well. This improves the comparability and stability of the estimated α^{gap} across time and portfolios.

Moreover, we can assign a positive alpha to the standardized factor, $\alpha^{gap} > 0$, and include this in our alpha model

$$\boldsymbol{\alpha}_t = \widehat{\boldsymbol{\alpha}}_t + \boldsymbol{x}_t^{gap} \alpha^{gap}. \tag{24}$$

This targets exactly the part of the trade that causes idiosyncratic drag when unclosed. Assigning a positive alpha to this gap encourages faster closure without interfering with factor moves.

We can estimate α^{gap} from data using the augmented Fama and MacBeth (1973) regression and calibrate its magnitude small relative to transaction costs to avoid over-penalizing.[2] In the absence of costs and constraints, $\boldsymbol{x}_t^{gap} = \boldsymbol{0}$ and the auxiliary factor has no effect.

As with any return factor, we include both the expected return and the risk of the gap factor when incorporating it into the optimization. The expected return is captured by α^{gap}, which is estimated from past returns. The risk is measured using the residual variance of the gap exposure, $(\boldsymbol{x}_t^{gap})' \boldsymbol{D}_t \boldsymbol{x}_t^{gap}$, consistent with the factor model's idiosyncratic risk metric. This ensures the optimizer accounts for both the performance drag and the uncertainty associated with holding residual exposures.

[2] Appendix B provides details for bounding the alpha for this proxy factor.

Optimization over Time

Although the auxiliary gap alpha α^{gap} may have limited impact in a single-period optimization, especially when transaction costs dominate, it plays a more important role over time. In a one-shot optimization, the optimizer generally still chooses to only partially close the trade gap $(\widehat{\boldsymbol{w}}t - \boldsymbol{w}t - 1)$ due to costs. The inclusion of $\boldsymbol{x}_t^{gap}\alpha^{gap}$ marginally encourages a move toward the target, but the result may not differ a great deal from the friction-aware solution without this term.

However, in a sequence of optimizations over time, this small incentive compounds. Without the gap alpha, the portfolio may drift away from the factor target $\widehat{\boldsymbol{w}}_t$ and remain there, as closing the gap is continually penalized by costs and receives no reward. With the gap alpha included, the optimizer sees a persistent positive incentive to reduce the residual gap, leading to a gradual reversion toward the target. This persistent âĂIJpullâĂİ can materially alter the portfolio trajectory over time, especially when the target moves gradually.

In this sense, the gap alpha introduces mean-reversion dynamics in residual space that are otherwise absent from standard optimization formulations. It makes the target not only an ideal but an attractor, guiding the optimized portfolio back toward factor space even when frictions create temporary deviations.

6.2 Liquidity-Sensitive Alpha Scaling

Alternatively, we can scale alphas for costly or illiquid names

$$\widetilde{\alpha}_{i,t} = s_i \widehat{\alpha}_{i,t}, \quad s_i > 1 \text{ for more costly/illiquid names.} \tag{25}$$

This pre-compensates for frictions where trading is slowest. We should use this cautiously. If no omitted factor is present, scaling can induce overtrading.

Other problems associated with inefficient alpha scaling may not manifest as negative idiosyncratic returns. Hentschel (2025a) describes how to scale alphas based on asset characteristics.

6.3 Idiosyncratic Risk Constraints and Penalties

In a fundamental factor model, a pure factor portfolio $\boldsymbol{w}^{(k)}$ is designed to have unit exposure to factor k and zero exposure to all other factors. This does not mean that the portfolio has no idiosyncratic (residual) risk. In fact, because the model's residual covariance matrix $\boldsymbol{D}_t$ is diagonal but nonzero, each pure factor portfolio has a positive residual variance

$$(\boldsymbol{w}^{(k)})'\boldsymbol{D}_t\boldsymbol{w}^{(k)} > 0. \tag{26}$$

Even if a portfolio is pure with respect to factors, it still carries stock-specific noise that the model's factors cannot explain, unless the factor portfolio holds a very large number of positions.

Raw Idiosyncratic Risk Constraints Distort Factor Exposures

A natural idea is to put a cap on the portfolio's idiosyncratic risk, for example

$$\boldsymbol{w}'\boldsymbol{D}_t\boldsymbol{w} \leq \sigma^2_{gap}. \tag{27}$$

This constraint limits the total residual standard deviation of the portfolio $\boldsymbol{w}$. But it applies equally to the built-in residual variance in the pure factor portfolios themselves, and any additional residual variance from deviations outside the factor space.

Since different pure factor portfolios have different levels of inherent idiosyncratic variance, often much higher for style factors such as value, size, or volatility than for market or sector factors, constraint equation (27) can create unintended effects. The optimizer is forced to tilt away from high-idiosyncratic-variance factors just to stay within the budget τ. The result is a portfolio whose factor mix is distorted relative to the intended factor composite $\widehat{\boldsymbol{w}}_t$.

A raw idiosyncratic risk limit doesn't distinguish between "good" idiosyncratic risk (the kind that's just part of the factor definitions) and "extra" idiosyncratic risk (from implementation frictions). As a result, it can unintentionally water down the factor bets we actually want.

6.4 Projection-Based Constraint on Idiosyncratic Risk

The solution is to measure idiosyncratic risk in the same metric in which the pure factor portfolios are residual-free. This is the metric used in the weighted least squares (WLS) regressions that define the pure factor returns in the risk model.

Let $\boldsymbol{\Gamma}$ be the symmetric positive-definite weighting matrix used in the factor-model regressions. We do not assume $\boldsymbol{\Gamma} = \boldsymbol{D}^{-1}$; $\boldsymbol{\Gamma}$ can reflect any reasonable scheme (residual-precision, liquidity, market-cap, or some hybrid). The key idea is to measure and constrain idiosyncratic exposure in the same metric that defines the factor space in estimation.

Define the $\boldsymbol{\Gamma}$-weighted projection onto the factor space spanned by $\boldsymbol{X}_t$:

$$\boldsymbol{P}_X = \boldsymbol{X}_t\big(\boldsymbol{X}_t'\boldsymbol{\Gamma}\boldsymbol{X}_t\big)^{-1}\boldsymbol{X}_t'\boldsymbol{\Gamma}, \qquad \boldsymbol{w}_t^{\perp} \equiv (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w}_t. \tag{28}$$

Here $\boldsymbol{w}_t^{\perp}$ is the part of the portfolio that lies outside factor space, defined consistently with the regression metric.

To measure idiosyncratic risk, use a positive semidefinite matrix $\dot{\boldsymbol{D}}_t$. The natural choice is $\dot{\boldsymbol{D}}_t = \boldsymbol{D}_t$, the residual covariance from the factor model. Then the residual (idiosyncratic) variance of $\boldsymbol{w}_t$ is

$$\mathrm{Var}_{\mathrm{id}}(\boldsymbol{w}_t) = (\boldsymbol{w}_t^{\perp})'\boldsymbol{D}_t\boldsymbol{w}_t^{\perp} = \boldsymbol{w}_t'(\boldsymbol{I} - \boldsymbol{P}_X)'\boldsymbol{D}_t(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w}_t. \tag{29}$$

More generally, if we prefer to budget residual risk in another metric (e.g., liquidity-scaled idiosyncratic risk), set $\dot{\boldsymbol{D}}_t \succcurlyeq 0$ and replace $\boldsymbol{D}_t$ by $\dot{\boldsymbol{D}}_t$ in equation (29). The projection-based idiosyncratic risk constraint is then

$$\boldsymbol{w}_t'(\boldsymbol{I} - \boldsymbol{P}_X)'\dot{\boldsymbol{D}}_t(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w}_t \leq \sigma_{gap}^2, \tag{30}$$

This limits idiosyncratic risk because it measures variance in the chosen risk metric.

Some optimizers may not accept such risk constraints using a second covariance matrix. In these cases, we can augment the regular covariance matrix with an additional projected idiosyncratic variance matrix and run the optimization. The augmented covariance matrix requires some recalibration of risk aversion or risk targets for the portfolio. The augmented optimization is

$$\max_{\boldsymbol{w}_t} \qquad \boldsymbol{w}_t'\boldsymbol{\alpha}_t - \frac{1}{2}\lambda\ \boldsymbol{w}_t'\widetilde{\boldsymbol{\Sigma}}_t\boldsymbol{w}_t \tag{31}$$

$$\widetilde{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t + \lambda_D(\boldsymbol{I} - \boldsymbol{P}_X)'\boldsymbol{D}_t(\boldsymbol{I} - \boldsymbol{P}_X), \tag{32}$$

where λ is the coefficient of risk aversion and $\lambda_D \geq 0$ controls the importance of the idiosyncratic risk. This covariance is similar to the augmented risk model of Saxena and Stubbs (2015) but respects the weighting method of factor regressions.

Any $\boldsymbol{w} \in col(\boldsymbol{X}_t)$ satisfies $(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{w} = \boldsymbol{0}$, so combinations of pure factor portfolios do not consume any part of the idiosyncratic risk budget in equation (30). The constraint or penalty acts only on the extra residual exposure beyond factor space, preventing penalties for pure factors with relatively high idiosyncratic risk, which occurs if we simply limit overall idiosyncratic risk.

The projection-based constraint equation (30) is the preferred method when the goal is to preserve the factor mix while limiting only the residual risk arising from implementation frictions. A constraint on raw idiosyncratic risk, like equation (27), mixes the inherent residual risk of the pure factor portfolios with the extra residual risk from true deviations. By contrast, equation (30) first removes the factor-space component (via $\boldsymbol{P}_X$) and only limits what remains. This preserves the intended factor mix while controlling the residual risk that arises from implementation frictions.

6.5 Combined Approach

The mean-variance optimization in equation (31) considers both expected returns and covariances. For return factors, we generally show both the expected returns and covariances. Here, for the omitted factors, we can do the same by using both the alpha updates from equation (24) and the augmented covariance from equation (32) in portfolio optimization. Neither the updated alpha nor the augmented covariance by itself are fair or sufficient corrections for the omitted factor problem.

7 Conclusion

When the target portfolio lies in factor space and residuals are iid mean-zero, optimized portfolios should have near-zero average idiosyncratic contributions. Persistent negative contributions imply a priced residual factor – an omitted factor in residual space.

Two simple tests, a HAC-robust mean test of portfolio residual returns and a Fama-MacBeth regression with projected gap exposures, differentiate bad luck from a genuine omitted factor.

If evidence supports the omitted factor view, remedies include liquidity-sensitive alpha scaling, adding an alpha gap factor, and using constraints or penalties on projection-based idiosyncratic risk.[3]

Reducing the impact of negative idiosyncratic returns is likely to increase transaction costs. A good optimization includes realistic performance measures for idiosyncratic returns, via expected returns and risks, in addition to the transaction costs. A balanced presentation is more likely to lead to a truly optimal portfolio.

[3] Appendix A and appendix B provide further measurement and calibration details.

8 References

Almgren, Robert, 2003, Optimal execution with nonlinear impact functions and trading-enhanced risk, *Applied Mathematical Finance* 10, 1–18.

Almgren, Robert, and Neil Chriss, 2001, Optimal execution of portfolio transactions, *Journal of Risk* 3, 5–39.

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

Hentschel, Ludger, 2025a, Contextual alpha: Emphasizing forecasts where they work best, Working paper, Versor Investments, New York, NY.

Hentschel, Ludger, 2025b, Preserving relative factor weights during optimization, Working paper, Versor Investments, New York, NY.

Markowitz, Harry, 1952, Portfolio selection, *Journal of Finance* 7, 77–91.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Saxena, Anureet, and Robert A. Stubbs, 2015, Augmented risk models to mitigate factor alignment problems, *Journal of Investment Management* 13, 57–79.

Sharpe, William F., 1974, Imputing expected returns from portfolio composition, *Journal of Financial and Quantitative Analysis* 9, 463–472.

A Net Return Gap Diagnostic

The net return gap is a practical diagnostic for gauging whether the return shortfall from deviating from the frictionless target $\widehat{\boldsymbol{w}}_t$ is broadly in line with measured transaction cost savings. It does not prove or refute the omitted-factor interpretation, but can help localize issues (e.g., cost-model miscalibration, constraint bottlenecks).

We define the idiosyncratic return gap (target minus actual) as

$$\Delta R^{id}_{t+1} = (\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_t)' \boldsymbol{\varepsilon}_{t+1}, \tag{33}$$

and the net friction-adjusted version as

$$\Delta R^{net}_{t+1} = (\boldsymbol{w}_t - \widehat{\boldsymbol{w}}_t)' \boldsymbol{\varepsilon}_{t+1} + \left(TC_{t+1} - TC^{target}_{t+1}\right), \tag{34}$$

where $TC_{t+1} = cost(\boldsymbol{w}_t - \boldsymbol{w}_{t-1})$ is the cost actually incurred and $TC^{target}_{t+1} = cost(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1})$ is the hypothetical cost of fully reaching the target. Thus, ΔR^{net}_{t+1} measures the net (return minus cost) benefit of implementing $\boldsymbol{w}_t$ instead of $\widehat{\boldsymbol{w}}_t$ over $[t, t+1]$.

Values near zero suggest a cost/return balance; persistently negative values imply return shortfall not justified by cost savings; persistently positive values suggest that costs are understated or that the optimizer is over-trading.

A.1 Estimation and Inference

We estimate the time-series mean

$$\overline{\Delta R}^{net} = \frac{1}{T} \sum_{t=0}^{T-1} \Delta R^{net}_{t+1} \tag{35}$$

and test $H_0 : E[\Delta R^{net}_{t+1}] = 0$ using a NeweyÐWest standard error matched to the trading horizon. As with idiosyncratic returns, winsorizing ΔR^{net}_{t+1} before estimation can improve robustness.

A balanced or small mean net gap does not guarantee that no omitted factor existsÑit only says the realized shortfall matched measured costs. The omitted-factor tests in the main text remain the primary tool for determining whether residual exposures have nonzero expected returns.

B Upper Bound on Gap-Factor Alpha

We sometimes consider assigning an alpha to the pure gap factor

$$\boldsymbol{x}^{gap}_t = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{g}_t, \qquad \boldsymbol{g}_t \equiv \widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}, \tag{36}$$

where $\boldsymbol{P}_X$ is the (weighted) projection matrix onto the factor space of the risk model and $\boldsymbol{g}_t$ is the trade required to move from the current portfolio $\boldsymbol{w}_{t-1}$ to the frictionless target $\widehat{\boldsymbol{w}}_t$.

By construction, $\boldsymbol{x}_t^{gap}$ lies entirely in the residual space. A positive alpha $\alpha^{gap} > 0$ attached to this factor encourages the optimizer to close residual-space gaps, pulling the solution toward the target.

We seek an upper bound (larger than we would use in practice) for α^{gap}: the value that would move the optimizer exactly back to the target portfolio in a single rebalancing step, assuming no other constraints bind. This is an extreme case – in reality, such a strong incentive would cause over-trading.

However, the bound is diagnostically valuable if the historically estimated α^{gap} for the gap factor is close to the bound, the optimizer is behaving as if closing the gap is extremely costly. If the estimate is far from the bound, there is room to make the alpha more positive to reduce persistent residual exposures without overshooting. Bounds are also valuable in many numerical searches, in case we wish to tune or optimize the gap alpha numerically.

B.1 Quadratic Costs

Suppose the optimizer solves

$$\max_{\boldsymbol{w}_t} \quad \boldsymbol{w}_t'(\widehat{\boldsymbol{\alpha}}_t + \alpha^{gap}\boldsymbol{x}_t^{gap}) - \frac{\lambda}{2}\boldsymbol{w}_t'\boldsymbol{\Sigma}\boldsymbol{w}_t - \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}_{t-1})'\boldsymbol{Q}(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}), \tag{37}$$

where $\widehat{\boldsymbol{\alpha}}_t$ are the implied alphas from the target portfolio, $\boldsymbol{\Sigma}$ is the factor-model covariance matrix, $\boldsymbol{Q} \succeq 0$ encodes quadratic trading costs, and $\lambda > 0$ is the risk-aversion parameter. The quadratic trading costs follow Almgren and Chriss (2001) and Almgren (2003).

If $\alpha^{gap} = 0$, the first-order condition is

$$\lambda\boldsymbol{\Sigma}\boldsymbol{w}_t + \boldsymbol{Q}(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) = \widehat{\boldsymbol{\alpha}}_t. \tag{38}$$

Since $\widehat{\boldsymbol{\alpha}}_t = \lambda\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_t$, this yields

$$\boldsymbol{w}_t = (\lambda\boldsymbol{\Sigma} + \boldsymbol{Q})^{-1}\big(\lambda\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_t + \boldsymbol{Q}\boldsymbol{w}_{t-1}\big). \tag{39}$$

With $\alpha^{gap} \neq 0$, the FOC becomes

$$\lambda\boldsymbol{\Sigma}\boldsymbol{w}_t + \boldsymbol{Q}(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) = \lambda\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_t + \alpha^{gap}\boldsymbol{x}_t^{gap}. \tag{40}$$

Requiring $\boldsymbol{w}_t = \widehat{\boldsymbol{w}}_t$ gives

$$\boldsymbol{Q}(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}) = \alpha^{gap}\boldsymbol{x}_t^{gap}, \tag{41}$$

or

$$\boldsymbol{Q}\boldsymbol{g}_t = \alpha^{gap}\boldsymbol{x}_t^{gap}. \tag{42}$$

Premultiplying by $(\boldsymbol{x}_t^{gap})'$ and using $\boldsymbol{x}_t^{gap} \neq \boldsymbol{0}$ yields the exact alpha that closes the gap in one step

$$\alpha_{\max}^{gap} = \frac{(\boldsymbol{x}_t^{gap})'\boldsymbol{Q}\boldsymbol{g}_t}{(\boldsymbol{x}_t^{gap})'\boldsymbol{x}_t^{gap}}. \tag{43}$$

If $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ and $\boldsymbol{Q} = \mathrm{diag}(q_i)$, the optimal frictionless trade toward the target in one period without a gap penalty is

$$w_{i,t} - w_{i,t-1} = \frac{\lambda\sigma_i^2}{\lambda\sigma_i^2 + q_i}\big(\widehat{w}_{i,t} - w_{i,t-1}\big). \tag{44}$$

To close the full gap in one step, the gap alpha must satisfy

$$\alpha_{\max,i}^{gap} = q_i\frac{\widehat{w}_{i,t} - w_{i,t-1}}{x_{i,t}^{gap}}. \tag{45}$$

This holds name-by-name; in practice, we would calibrate α^{gap} from the portfolio-level projection in (43).

B.2 Linear Costs

If transaction costs are proportion to bid-ask spreads, the costs are proportional to trade size, $TC(\Delta\boldsymbol{w}) = \boldsymbol{c}'|\Delta\boldsymbol{w}|$. For general $\boldsymbol{\Sigma}$, the optimality condition produces a polyhedral no-trade region

$$\big|\lambda(\boldsymbol{\Sigma}(\boldsymbol{w}_t - \widehat{\boldsymbol{w}}_t))_i\big| \le c_i \quad \Rightarrow \quad w_{i,t} = w_{i,t-1}. \tag{46}$$

If $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$, the no-trade condition simplifies to

$$\big|\lambda\sigma_i^2(w_{i,t} - \widehat{w}_{i,t})\big| \le c_i. \tag{47}$$

The one-step bound on α^{gap} is simply the shift needed to push each i outside its no-trade band

$$\alpha_{\max,i}^{gap} = \mathrm{sign}(g_{i,t})c_i + \lambda\sigma_i^2 g_{i,t}. \tag{48}$$

B.3 Square-Root Impact Costs

For "square-root" impact, trading costs are $TC(\Delta w_i) = k_i|\Delta w_i|^{3/2}$, and marginal cost is proportional to $|\Delta w_i|^{1/2}$. The optimality condition for diagonal $\boldsymbol{\Sigma}$ is

$$\lambda\sigma_i^2(\widehat{w}_{i,t} - w_{i,t}) = k_i\,\mathrm{sign}(\Delta w_i)|\Delta w_i|^{1/2}. \tag{49}$$

Solving for Δw_i in closed form is straightforward

$$\Delta w_i = \left(\frac{\lambda\sigma_i^2}{k_i}\right)^2 \left(\widehat{w}_{i,t} - w_{i,t}\right)^2. \tag{50}$$

The one-step bound on α^{gap} is the alpha shift that induces exactly this Δw_i when the gap equals $\widehat{w}_{i,t} - w_{i,t-1}$.

Among these models, quadratic costs with general $\boldsymbol{\Sigma}$ yield the most tractable and interpretable upper bound for α^{gap}. The diagonal cases are algebraically simple and help illustrate the mechanics of the bound, even though real-world risk models have substantial off-diagonal structure. Although we cannot solve for the upper bound on alpha analytically, we can find this bound via numerical optimization.

C Augmented Covariance in Standard Factor Form

The augmented covariance matrix in equation (32) penalizes residual exposure through an additional variance term

$$\widetilde{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t + \lambda_D(\boldsymbol{I} - \boldsymbol{P}_X)'\boldsymbol{D}_t(\boldsymbol{I} - \boldsymbol{P}_X), \tag{51}$$

where $\boldsymbol{\Sigma}_t = \boldsymbol{X}_t\boldsymbol{\Omega}_t\boldsymbol{X}_t' + \boldsymbol{D}_t$ is the standard factor-model covariance matrix.

This augmented matrix does not initially appear to conform to the standard 3-part representation expected by many optimization platforms, which require

1. an $n \times K$ matrix of factor exposures $\boldsymbol{X}_t$,
2. a $K \times K$ factor covariance matrix $\boldsymbol{\Omega}_t$,
3. and an $n \times n$ diagonal matrix of idiosyncratic variances $\boldsymbol{D}_t$ (possibly delivered as an $n \times 1$ vector),

so that the software can construct

$$\boldsymbol{\Sigma}_t = \boldsymbol{X}_t\boldsymbol{\Omega}_t\boldsymbol{X}_t' + \boldsymbol{D}_t. \tag{52}$$

Representing a factor-based covariance matrix in its constituent form can provide large computational speed advantages in matrix inversion, especially when the covariance matrix is large.

The augmented covariance matrix $\widetilde{\boldsymbol{\Sigma}}_t$ can be rewritten in exactly this form by introducing an additional synthetic factor that captures the residual penalty.

C.1 Gap Direction in Residual Space

Let the gap vector be

$$\boldsymbol{x}_t^{gap} = (\boldsymbol{I} - \boldsymbol{P}_X)(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_{t-1}), \tag{53}$$

the part of the desired trade that lies in residual space. Define a unit-norm residual direction

$$\boldsymbol{z}_t \equiv \frac{\boldsymbol{x}_t^{gap}}{\|\boldsymbol{x}_t^{gap}\|}, \quad \text{so that} \quad \|\boldsymbol{z}_t\| = 1. \tag{54}$$

This synthetic factor has exposure only in residual space. Its risk contribution under $\boldsymbol{D}_t$ is

$$v_t^{gap} \equiv \boldsymbol{z}_t' \boldsymbol{D}_t \boldsymbol{z}_t. \tag{55}$$

C.2 Augmented Factor Model

We now define an augmented exposure matrix

$$\widetilde{\boldsymbol{X}}_t = [\boldsymbol{X}_t \quad \boldsymbol{z}_t], \tag{56}$$

which has dimension $n \times (K+1)$, and an augmented factor covariance matrix

$$\widetilde{\boldsymbol{\Omega}}_t = \begin{bmatrix} \boldsymbol{\Omega}_t & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_D v_t^{gap} \end{bmatrix}, \tag{57}$$

which has dimensions $(K+1) \times (K+1)$ and is positive semidefinite. As before, λ_D measures the importance of the additional risk and requires calibration.

Then the total covariance matrix can be written

$$\widetilde{\boldsymbol{\Sigma}}_t = \boldsymbol{X}_t \boldsymbol{\Omega}_t \boldsymbol{X}_t' + \boldsymbol{D}_t + \lambda_D v_t^{gap} \boldsymbol{z}_t \boldsymbol{z}_t' \tag{58}$$

$$= \widetilde{\boldsymbol{X}}_t \widetilde{\boldsymbol{\Omega}}_t \widetilde{\boldsymbol{X}}_t' + \boldsymbol{D}_t. \tag{59}$$

Thus, the augmented covariance matrix $\widetilde{\boldsymbol{\Sigma}}_t$ can be expressed in standard factor form with

1. an $n \times (K+1)$ matrix of factor exposures $\widetilde{\boldsymbol{X}}_t$,
2. a $(K+1) \times (K+1)$ factor covariance matrix $\widetilde{\boldsymbol{\Omega}}_t$,
3. and an $n \times n$ diagonal matrix of idiosyncratic variances $\boldsymbol{D}_t$ (possibly delivered as an $n \times 1$ vector).

This reformulation interprets the residual penalty as coming from a synthetic factor with exposure direction $\boldsymbol{z}_t$ (purely in residual space), estimated risk $\boldsymbol{z}_t'\boldsymbol{D}_t\boldsymbol{z}_t$, and subjective importance λ_D. Using the same projection matrix $\boldsymbol{P}_{\mathrm{X}}$ as in the main factor regressions ensures that the residual penalty applies only to directions orthogonal to $\boldsymbol{X}_t$ under the same weighting metric.

This transformation allows standard portfolio optimization software to implement the residual risk penalty by treating it as a regular return factor with known exposure and estimated variance.