Brinson and Factors: A Unified Framework for Complete Portfolio Return Attribution

Ludger Hentschel

November 2024

Abstract

Factor-based attribution leaves residuals, and Brinson-style attribution introduces an interaction term that many investors find uninformative. Both obscure the link between portfolio performance and the manager's actual investment decisions.

We show that these limitations can be resolved within a unified regression-based framework. By expressing both factor and Brinson attributions as weighted regressions and imposing a single linear restriction that enforces active-exactness, forcing the portfolio-weighted residual or interaction terms to vanish, we reallocate these nuisance components in the statistically least-distorting way. The resulting attributions are complete: every element of portfolio performance is explained by identifiable investment decisions, with adjustments proportional to their statistical uncertainty.

This approach combines the statistical rigor of factor attribution with the intuitive structure of Brinson analysis, yielding an internally consistent and decision-based foundation for complete portfolio return attribution. For practitioners, it produces cleaner reports, clearer interpretation, and a principled alternative to ad hoc reallocations of residuals and interaction effects.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

Acknowledgements
I am grateful for helpful comments from Tony Elavia, Harvey Fram, and Migene Kim.

1 Introduction

Portfolio return attribution is a central tool in performance evaluation for investment funds. The classical allocation-selection framework of Brinson and Fachler (1985) and Brinson, Hood, and Beebower (1986) decomposes active returns into allocation, selection, and interaction effects. Alternatively, factor-based performance attribution uses regression techniques to explain portfolio returns through exposures to systematic factors and residual idiosyncratic returns (see, for example, Grinold (2006)).

Spaulding (2003) argues that a complete attribution should satisfy two "laws": it must attribute performance to the manager's actual decisions and exhaust all returns without residuals. We formalize these principles as linear restrictions within a regression framework, ensuring that all portfolio returns are assigned to identifiable investment decisions in a statistically efficient and internally consistent way.

Standard factor attribution separates portfolio returns into factor components and unidentified residuals. These residuals are an inevitable part of factor regressions but represent a nuisance for performance attribution. We reallocate these residuals to factor returns using a restricted generalized least squares (GLS) update that enforces completeness while minimizing statistical distortion. Because the restriction adds only one degree of constraint, the adjustments are typically small and fall on the least precisely estimated factor returns.

The same logic extends to allocation-selection attribution, where the interaction term plays a role analogous to residuals in factor models. In Brinson, Hood, and Beebower (1986), portfolio returns are decomposed into allocation across groups, selection within groups, and an interaction term measuring how well selection worked in overweighted groups. This interaction is usually treated as a nuisance and reallocated to allocation or selection components through ad hoc rules.[1] Although Brinson-style attribution is commonly treated as a calculation, we reinterpret this framework statistically, allowing interaction effects to be redistributed across allocation and selection components according to their relative estimation precision, in a way that minimizes the statistical distortion to both.

Practical guides such as Bacon (2004) have made these attribution approaches mainstream, but they leave either residuals or interaction terms that can obscure the manager's true contribution. We unify these approaches

[1] Spaulding (2008) describes some common static reallocation rules and proposes a set of conditional reallocation rules.

within a single regression-based framework that provides a principled statistical foundation for reallocating nuisance components.

Section 2 introduces the restricted-regression method for reallocating residual return components in factor attribution. Section 3 derives the Brinson, Hood, and Beebower (1986) allocation-selection decomposition from benchmark- and portfolio-weighted regressions and applies the same restriction to eliminate interaction effects. Section 4 presents illustrative examples showing how the reallocation depends on relative estimation precision. Section 5 concludes.

2 Factor-Based Attribution

Suppose we have N assets and write the N-dimensional vectors $\boldsymbol{r}$ for returns, $\boldsymbol{w}_P$ for portfolio weights, and $\boldsymbol{w}_B$ for benchmark weights. Let the $N \times K$ matrix $\boldsymbol{X}$ contain the assets' exposures to K factors, which may include both risk factors and alpha factors. (See Connor (1995) and Grinold (2006), for example.) In broader practice, such factor structures often trace back to empirical asset pricing models such as Fama and French (1993). Throughout, bold symbols denote vectors and matrices; scalars are not bold.

2.1 Classical decomposition

Factor attribution starts from the cross-sectional linear model for returns

$$\boldsymbol{r} = \boldsymbol{X}\boldsymbol{f} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{f}$ are factor returns and $\boldsymbol{u}$ are idiosyncratic (security-specific) residuals. Asset returns, factor exposures, factor returns, and residual returns vary over time. For simplicity, we suppress time subscripts on all variables.

The portfolio's active return decomposes as

$$\begin{aligned} r_A &= \boldsymbol{w}_P'\boldsymbol{r} - \boldsymbol{w}_B'\boldsymbol{r} \\ &= (\boldsymbol{w}_P - \boldsymbol{w}_B)'\boldsymbol{X}\boldsymbol{f} + (\boldsymbol{w}_P - \boldsymbol{w}_B)'\boldsymbol{u} \\ &= \boldsymbol{w}_A'\boldsymbol{X}\boldsymbol{f} + \boldsymbol{w}_A'\boldsymbol{u}, \end{aligned} \tag{2}$$

where $\boldsymbol{w}_A \equiv \boldsymbol{w}_P - \boldsymbol{w}_B$ are the active portfolio weights.

The first term is the factor-driven return contribution. The second term is the unexplained residual return contribution.

If the factors are generic risk factors, investors often refer to the residual contribution as "security selection." While this may be technically correct, it obscures essentially all investment decisions made by the portfolio manager.

A clearer approach includes alpha factors in $\boldsymbol{X}$ and $\boldsymbol{f}$ that drive the main investment decisions. The return contributions from these alpha factors lay bare the success or failures of the individual investment decisions. In this case, however, interpreting the residual return contributions as "security selection" is no longer correct. The residual returns now capture deviations from the factors. Such deviations may arise because the factor model does not fully capture all of the factors or because the portfolio construction leads to deviations from the alpha factor portfolios.

Especially in systematic investment processes, the portfolio manager generally tries to minimize deviations from the alpha factor portfolios. As a result, we should expect the residual return contributions to be relatively small. Even when they are small, however, the residual return contributions can be a nuisance in performance reporting.

2.2 Regression formulation

We generally estimate factor returns via a cross-sectional regression, possibly with a weight matrix $\boldsymbol{W}_f$, which yields the factor return estimates

$$\widehat{\boldsymbol{f}} = (\boldsymbol{X}'\boldsymbol{W}_f\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_f\boldsymbol{r}. \tag{3}$$

Common choices for the weighting matrix are $\boldsymbol{W}_f = \boldsymbol{I}$ (for OLS), $\boldsymbol{W}_f = \text{diag}(\text{market cap})$ (for market cap weighting), or $\boldsymbol{W}_f = \text{diag}(\text{precision})$ (for inverse-variance weighting). Different $\boldsymbol{W}_f$ reflect different views about which securities should anchor the cross-sectional "factor plane." Our qualitative results are not sensitive to reasonable choices of $\boldsymbol{W}_f$.

2.3 Reallocating residuals

To avoid a residual line item at the portfolio level, we can estimate restricted factor returns $\widehat{\boldsymbol{f}}^*$ that satisfy

$$\boldsymbol{w}_A'\boldsymbol{X}\widehat{\boldsymbol{f}}^* = \boldsymbol{w}_A'\boldsymbol{r}. \tag{4}$$

This restriction implies

$$\boldsymbol{w}_A'\boldsymbol{u} = 0, \tag{5}$$

so that we reallocate what would otherwise appear as a residual return contribution to the factor return contributions.

Although Greene and Seaks (1991) argue that solving the associated Lagrangian system is numerically preferable we also present the analytical

solution because it shows exactly how the restriction reallocates residuals across factors.

Analytical solution

In our case, we estimate the factor returns using GLS with weighting matrix $\boldsymbol{W}_f$ but impose a linear constraint using the unweighted, raw asset returns $\boldsymbol{r}$. We focus on the raw returns since they are used in attribution.

Define

$$\boldsymbol{\Omega} = (\boldsymbol{X}'\boldsymbol{W}_f\boldsymbol{X})^{-1}. \tag{6}$$

to be the usual covariance of the estimated factor returns.

Theil (1971) derives the restricted GLS estimate of the factor returns, which we can write as a single-constraint update of the unconstrained factor return estimates $\widehat{\boldsymbol{f}}$,

$$\widehat{\boldsymbol{f}}^* = \widehat{\boldsymbol{f}} + \boldsymbol{\Omega}\,\boldsymbol{X}'\boldsymbol{w}_A\left(\boldsymbol{w}_A'\boldsymbol{X}\,\boldsymbol{\Omega}\,\boldsymbol{X}'\boldsymbol{w}_A\right)^{-1}\boldsymbol{w}_A'(\boldsymbol{r} - \boldsymbol{X}\widehat{\boldsymbol{f}}). \tag{7}$$

This is a textbook result for GLS estimates subject to linear restrictions on the coefficients.[2]

This solution adjusts the unconstrained factor returns $\widehat{\boldsymbol{f}}$ just enough to satisfy the restriction $\boldsymbol{w}_A'(\boldsymbol{r} - \boldsymbol{X}\widehat{\boldsymbol{f}}^*) = 0$, so that the portfolio-weighted residuals vanish and the active portfolio return is fully explained by the factor contributions,

$$r_A = \boldsymbol{w}_A'\boldsymbol{X}\widehat{\boldsymbol{f}}^*. \tag{8}$$

The overall adjustment is proportional to $\boldsymbol{\Omega}$, so the restricted estimate makes larger changes to those factor returns that are estimated with greater uncertainty. Sneddon (2021) highlights that factor attribution commonly contains estimation noise. Here, we allocate the residual return contributions to the factor return contributions in a manner that takes the best advantage of the flexibility offered by estimation noise.

The adjustment term $\boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{w}_A\left(\boldsymbol{w}_A'\boldsymbol{X}\boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{w}_A\right)^{-1}$ can be interpreted as the regression coefficient of the factor vector on the fitted active return. Since $\boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{w}_A = \mathrm{Cov}(\widehat{\boldsymbol{f}}, \widehat{r}_A)$ and $\boldsymbol{w}_A'\boldsymbol{X}\boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{w}_A = \mathrm{Var}(\widehat{r}_A)$, the update $\widehat{\boldsymbol{f}}^* -$

[2] See, for example, equation (8.5) in Theil (1971). Let the restriction be $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$ with $\boldsymbol{R} \equiv \boldsymbol{w}_A'\boldsymbol{X}$ and $\boldsymbol{q} \equiv \boldsymbol{w}_A'\boldsymbol{r}$. Theil's formula is $\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_\beta\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{\Sigma}_\beta\boldsymbol{R}')^{-1}(\boldsymbol{q} - \boldsymbol{R}\widehat{\boldsymbol{\beta}})$. After substituting $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{f}}$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Omega}$, we find $\widehat{\boldsymbol{f}}^* = \widehat{\boldsymbol{f}} + \boldsymbol{\Omega}\,\boldsymbol{X}'\boldsymbol{w}_A\left(\boldsymbol{w}_A'\boldsymbol{X}\,\boldsymbol{\Omega}\,\boldsymbol{X}'\boldsymbol{w}_A\right)^{-1}\boldsymbol{w}_A'(\boldsymbol{r} - \boldsymbol{X}\widehat{\boldsymbol{f}})$. Alternatively, the result follows from the more common textbook form of the restricted OLS estimator after replacing the covariance of the unrestricted estimates, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, with the GLS covariance, $\boldsymbol{\Omega}$. See Greene and Seaks (1991), for example.

$\widehat{\boldsymbol{f}} = \beta_{f|\hat{r}_A}\, \boldsymbol{w}_A'(\boldsymbol{r} - \boldsymbol{X}\widehat{\boldsymbol{f}})$ adjusts factor estimates in proportion to their beta with the fitted active return and in proportion to the distance from the constraint. Factors that are more relevant, those with higher beta to the active fitted return, absorb a larger share of the adjustment, and factors with larger estimation variance move farther because such shifts incur a smaller statistical cost.

In statistical terms, the restriction redistributes the unexplained returns across factors in the least-distorting way, satisfying the active-exactness constraint with minimal loss of likelihood.

Discussion

The restriction ensures that the portfolio has no residual return contributions. This can be extremely convenient when the factors include the primary alpha factors that drive the portfolio weights. Although the portfolio most likely incurred some residual returns, these returns are a reporting nuisance, even when they represent a small part of the total returns.

Investors often use ad hoc methods to distribute the residual returns to the factor returns in order to avoid itemizing them.

The restricted regression statistically reallocates the residual returns to the factor returns in a way that has the smallest impact on the overall fit of the factor regression. When the estimation universe contains a substantial number of securities N, this single linear restriction generally has minimal impact on the regression fit. The reallocation mostly changes the factor returns that are estimated with the least precision.

It is important, however, that the portfolio manager remains aware of the residual returns. Especially when residual returns are material, they may warrant the attention of the portfolio manager. For internal analysis, it probably is wise to consider both the unrestricted attribution and the restricted attribution.

Next, we apply the same logic to Brinson-style attribution: We interpret the attribution in regression terms and impose a single linear restriction in order to remove the interaction term by the smallest statistically justified adjustment.

3 Brinson Attribution

Brinson and Fachler (1985) and Brinson, Hood, and Beebower (1986) describe attribution that partitions active returns by groups (e.g., sectors). Each asset belongs to a group $g(i) \in 1, \ldots, G$, and the attribution divides portfolio return contributions into allocation decisions across groups, selec-

tion decisions within groups, and an interaction term that ensures that the decomposition exactly matches the portfolio's active return.

3.1 Classical decomposition

The classical Brinson decomposition of the active return r_A is

$$\begin{aligned} r_A = & \sum_g (w_{P,g} - w_{B,g})(\bar{r}_{B,g} - r_B) + \\ & \sum_g w_{B,g}(\bar{r}_{P,g} - \bar{r}_{B,g}) + \\ & \sum_g (w_{P,g} - w_{B,g})(\bar{r}_{P,g} - \bar{r}_{B,g}). \end{aligned} \tag{9}$$

Here, $w_{P,g}$ and $w_{B,g}$ are the portfolio and benchmark exposures to group g, $\bar{r}_{P,g}$ and $\bar{r}_{B,g}$ are the corresponding within-group returns, and r_B is the total benchmark return.

The three terms represent allocation, selection, and interaction effects, respectively. The interaction term has a natural covariance interpretation: It is positive when the portfolio has positive allocation effects in groups where it also has positive selection effects. Nonetheless, the interaction term is generally treated as a nuisance: necessary for arithmetic completeness but not easily interpreted on its own. As a result, investors commonly reallocate it to the allocation or selection components using ad hoc rules.

3.2 Regression formulation

The attribution in equation (9) is usually interpreted as a mechanical arithmetic identity. But this view obscures the structure that makes the Brinson results consistent with standard regression-based attributions. We can express the same relationships in regression form without changing any of the components. Doing so clarifies the estimation precision of the allocation and selection terms and provides a statistical foundation for a principled reallocation of the interaction term. This may seem like a conceptual leap from the Brinson arithmetic, but it relies only on standard results from weighted least squares regression.

We can view the within-group averages $\bar{r}_{P,g}$ and $\bar{r}_{B,g}$ as estimates of group mean returns obtained from two weighted least squares (WLS) regressions. Let $\boldsymbol{D}$ be an $(N \times G)$ dummy matrix indicating each asset's group membership. The benchmark- and portfolio-weighted regressions are

$$\boldsymbol{r} = \boldsymbol{D}\boldsymbol{\mu}_B + \boldsymbol{\varepsilon}_B, \qquad \widehat{\boldsymbol{\mu}}_B = (\boldsymbol{D}'\boldsymbol{W}_B\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}_B\boldsymbol{r}, \tag{10}$$

and

$$\boldsymbol{r} = \boldsymbol{D}\boldsymbol{\mu}_P + \boldsymbol{\varepsilon}_P, \qquad \widehat{\boldsymbol{\mu}}_P = (\boldsymbol{D}'\boldsymbol{W}_P\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}_P\boldsymbol{r}, \tag{11}$$

with $\boldsymbol{W}_B = \text{diag}(\boldsymbol{w}_B)$ and $\boldsymbol{W}_P = \text{diag}(\boldsymbol{w}_P)$. Each regression simply computes the weighted average return in each group. This is exactly what the classical Brinson arithmetic does.

These regressions correspond to the unique factor models that reproduce Brinson-style attribution exactly. Each model contains only the group-dummy factors, and the weighting matrices must be the benchmark and portfolio weights, $\boldsymbol{W}_B$ and $\boldsymbol{W}_P$, respectively. Using any other factors or weights would produce a different decomposition and therefore depart from the classical Brinson results. In contrast, elsewhere in this paper our generalized GLS framework allows flexible factor sets and weighting matrices. Here, however, the dummy-only specification with these specific weights is the only one consistent with the standard Brinson components.

Thus, the fitted means $\widehat{\mu}_{B,g} = (\sum_{i\in g} w_{B,i}r_i)/(\sum_{i\in g} w_{B,i})$ and $\widehat{\mu}_{P,g} = (\sum_{i\in g} w_{P,i}r_i)/(\sum_{i\in g} w_{P,i}) = \overline{r}_{P,g}$ are the familiar group average returns for the benchmark and portfolio, respectively. The fitted values

$$\widehat{\boldsymbol{r}}_B = \boldsymbol{D}\widehat{\boldsymbol{\mu}}_B, \qquad \widehat{\boldsymbol{r}}_P = \boldsymbol{D}\widehat{\boldsymbol{\mu}}_P, \tag{12}$$

reproduce the benchmark's and portfolio's group-level fitted returns. We then have $\widehat{r}_{B,g} = \overline{r}_{B,g}$ and $\widehat{r}_{P,g} = \overline{r}_{P,g}$. To emphasize the regression source of these values, we will use the "hat" notation for these fitted returns in the remainder.

The difference between the fitted returns, $\widehat{\boldsymbol{r}}_P - \widehat{\boldsymbol{r}}_B = \boldsymbol{D}(\widehat{\boldsymbol{\mu}}_P - \widehat{\boldsymbol{\mu}}_B)$, represents the change in estimated group means when moving from benchmark to portfolio weighting.

Expressed in this notation, the Brinson decomposition is

$$\begin{aligned} r_A \;&= \boldsymbol{w}_A'(\widehat{\boldsymbol{r}}_B - r_B\boldsymbol{\iota}) + \\ &\quad\; \boldsymbol{w}_B'(\widehat{\boldsymbol{r}}_P - \widehat{\boldsymbol{r}}_B) + \\ &\quad\; \boldsymbol{w}_A'(\widehat{\boldsymbol{r}}_P - \widehat{\boldsymbol{r}}_B), \end{aligned} \tag{13}$$

where $\boldsymbol{\iota}$ is an N-vector of ones and $\boldsymbol{w}_A = \boldsymbol{w}_P - \boldsymbol{w}_B$. The three terms correspond exactly to allocation, selection, and interaction effects, as in the arithmetic decomposition of equation (9). The regression representation simply provides a statistical framework for reasoning about their sampling variability, covariance, and optimal reallocation.

Having cast Brinson attribution as two parallel WLS regressions, we now apply the same single active-exactness restriction used for factors (section 2.3) to the stacked group-mean estimates. Imposing $\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\boldsymbol{m} = 0$ eliminates the interaction term and, via restricted GLS, adjusts $\widehat{\boldsymbol{\mu}}_B$ and $\widehat{\boldsymbol{\mu}}_P$ by the smallest GLS distance; see section 3.4.

3.3 Reallocating interaction effects

Rather than ad hoc splits, the restricted-GLS update redistributes the interaction across allocation and selection according to their relative imprecision encoded in $\boldsymbol{\Gamma}$: precise components move little, noisier ones move more. Equivalently, the share absorbed by each side is proportional to its covariance (beta) with the fitted active return (see section 3.4). When the two regressions have similar precision, the split is near 50/50; if the portfolio-weighted regression is less precise (e.g., because of greater portfolio concentration), more of the interaction is assigned to allocation, and conversely when it is more precise.

3.4 Active-exactness as a restriction on both regressions

Our goal is to reallocate the interaction term by imposing a single linear restriction that makes the active return equal the sum of allocation and selection effects. We seek adjusted group means $\widehat{\boldsymbol{\mu}}_B^*$ and $\widehat{\boldsymbol{\mu}}_P^*$ such that, using the active portfolio weights $\boldsymbol{w}_A$, the portfolio-weighted difference in group means is zero

$$\boldsymbol{w}_A' \boldsymbol{D} \left(\widehat{\boldsymbol{\mu}}_P^* - \widehat{\boldsymbol{\mu}}_B^* \right) = 0. \tag{14}$$

This "active-exactness" restriction ensures that the total active return r_A is fully explained by the adjusted allocation and selection effects, with no residual or interaction term.

To express this compactly, define the stacked vector of group-mean estimates

$$\widehat{\boldsymbol{m}} \equiv \begin{bmatrix} \widehat{\boldsymbol{\mu}}_B \\ \widehat{\boldsymbol{\mu}}_P \end{bmatrix}, \tag{15}$$

and let $\boldsymbol{H}$ extract the difference between the two halves,

$$\boldsymbol{H} \equiv \begin{bmatrix} -\boldsymbol{I}_G & \boldsymbol{I}_G \end{bmatrix}, \qquad \text{so that} \quad \boldsymbol{H}\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{\mu}}_P - \widehat{\boldsymbol{\mu}}_B. \tag{16}$$

Then the active-exactness condition in equation (14) can be written as a single linear restriction on the stacked means

$$\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\widehat{\boldsymbol{m}} = 0. \tag{17}$$

Among all adjusted means $\boldsymbol{m}^*$ that satisfy this restriction, we want the one that is statistically most plausible, the one closest to the original estimates in the GLS sense. This leads to the restricted GLS problem

$$\min_{\boldsymbol{m}} \ \tfrac{1}{2}(\boldsymbol{m} - \widehat{\boldsymbol{m}})' \boldsymbol{\Gamma}^{-1} (\boldsymbol{m} - \widehat{\boldsymbol{m}}) \quad \text{s.t.} \quad \boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\boldsymbol{m} = 0, \tag{18}$$

where $\boldsymbol{\Gamma}$ is the joint covariance of the stacked estimates,

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Sigma}_B & \boldsymbol{\Sigma}_{B,P} \\ \boldsymbol{\Sigma}_{P,B} & \boldsymbol{\Sigma}_P \end{bmatrix}. \tag{19}$$

This is a standard GLS problem with a single linear restriction. As for the restricted factor returns, we can follow Theil (1971) and write the restricted estimate of the group means as

$$\widehat{\boldsymbol{m}}^* = \widehat{\boldsymbol{m}} + \boldsymbol{\Gamma}\,\boldsymbol{H}'\boldsymbol{D}'\boldsymbol{w}_A \left(\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\,\boldsymbol{\Gamma}\,\boldsymbol{H}'\boldsymbol{D}'\boldsymbol{w}_A\right)^{-1} \left(0 - \boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\widehat{\boldsymbol{m}}\right). \tag{20}$$

This makes small adjustments to the group means in $\widehat{\boldsymbol{\mu}}_B$ and $\widehat{\boldsymbol{\mu}}_P$ in order to just satisfy the constraint that the interaction term in the attribution is 0.

The size of the adjustment depends on two components of this expression: the covariance matrix $\boldsymbol{\Gamma}$, which captures the statistical uncertainty of the group-mean estimates, and the deviation from the restriction, $(0 - \boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\widehat{\boldsymbol{m}})$, which measures how far the unrestricted estimates violate active-exactness. The covariance governs the relative size of the adjustments. Group means that are more uncertain or more inconsistent with the restriction receive proportionally larger adjustments, while precise or already-consistent estimates remain largely unchanged. The deviation from the restriction governs the total size of the adjustment.

Unstacking $\widehat{\boldsymbol{m}}^*$ into its benchmark and portfolio components yields the adjusted group means

$$\widehat{\boldsymbol{m}}^* = \begin{bmatrix} \widehat{\boldsymbol{\mu}}_B^* \\ \widehat{\boldsymbol{\mu}}_P^* \end{bmatrix}, \qquad \text{so that} \quad \widehat{\boldsymbol{r}}_B^* = \boldsymbol{D}\widehat{\boldsymbol{\mu}}_B^*, \quad \widehat{\boldsymbol{r}}_P^* = \boldsymbol{D}\widehat{\boldsymbol{\mu}}_P^*. \tag{21}$$

By construction, the restriction $\boldsymbol{w}_A' \boldsymbol{D}(\widehat{\boldsymbol{\mu}}_P^* - \widehat{\boldsymbol{\mu}}_B^*) = 0$ is exactly satisfied, so that the interaction component of the Brinson decomposition vanishes while

both the benchmark- and portfolio-weighted regressions remain as close as possible (in GLS distance) to their unconstrained estimates.

With these adjusted fitted vectors, the complete attribution without an interaction term is

$$r_A = A^* + S^*, \qquad A^* = \boldsymbol{w}_A'(\widehat{\boldsymbol{r}}_B^* - r_B\boldsymbol{\iota}), \qquad S^* = \boldsymbol{w}_B'(\widehat{\boldsymbol{r}}_P^* - \widehat{\boldsymbol{r}}_B^*). \tag{22}$$

The constraint ensures that the total active return equals the sum of allocation and selection effects, while the GLS weighting ensures that the adjustments to $\widehat{\boldsymbol{r}}_B$ and $\widehat{\boldsymbol{r}}_P$ are statistically optimal – larger for the noisier estimates, smaller for the precise ones.

Precision inputs for the stacked restriction

When a reliable full return covariance $\boldsymbol{\Sigma}$ is available, for example from a risk model, we obtain the regression residual covariance under generic weights $\boldsymbol{W}$ from the WLS hat matrix in raw return space,

$$\boldsymbol{P}_W = \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}, \qquad \boldsymbol{M}_W = \boldsymbol{I} - \boldsymbol{P}_W, \qquad \boldsymbol{\Delta} = \boldsymbol{M}_W\,\boldsymbol{\Sigma}\,\boldsymbol{M}_W'. \tag{23}$$

In practice we set $\boldsymbol{W} = \boldsymbol{W}_B$ for consistency with benchmark-relative reporting, but any reasonable choice (cap or precision weights) only affects precision weighting, not the active-exactness identity.[3]

Because the estimates $\widehat{\boldsymbol{\mu}}_B$ and $\widehat{\boldsymbol{\mu}}_P$ are linear in $\boldsymbol{r}$, we write

$$\widehat{\boldsymbol{\mu}}_B = D'\boldsymbol{A}_B\boldsymbol{r}, \qquad D'\boldsymbol{A}_B \equiv (\boldsymbol{D}'\boldsymbol{W}_B\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}_B, \tag{24}$$

$$\widehat{\boldsymbol{\mu}}_P = D'\boldsymbol{A}_P\boldsymbol{r}, \qquad D'\boldsymbol{A}_P \equiv (\boldsymbol{D}'\boldsymbol{W}_P\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}_P. \tag{25}$$

and obtain the needed covariances by propagation

$$\boldsymbol{\Sigma}_B = D'\boldsymbol{A}_B\,\boldsymbol{\Delta}\,D'\boldsymbol{A}_B', \quad \boldsymbol{\Sigma}_P = D'\boldsymbol{A}_P\,\boldsymbol{\Delta}\,D'\boldsymbol{A}_P', \quad \boldsymbol{\Sigma}_{B,P} = D'\boldsymbol{A}_B\,\boldsymbol{\Delta}\,D'\boldsymbol{A}_P'. \tag{26}$$

These blocks define $\boldsymbol{\Gamma}$ used in the restricted GLS update (20).

If $\boldsymbol{\Sigma}$ is unavailable, we can use a shrinkage estimator of the sample covariance, as suggested by Ledoit and Wolf (2004). As fallbacks, the diagonal sample variances $\boldsymbol{\Delta} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_N^2)$ or even $\boldsymbol{\Delta} = \boldsymbol{I}$ are acceptable. These only affect the precision weights. The constraint guarantees exactness under any reasonable covariance matrix.

[3] In principle, one could compute separate residual covariances $\boldsymbol{\Delta}_B = \boldsymbol{M}_{W_B}\boldsymbol{\Sigma}\boldsymbol{M}_{W_B}'$ and $\boldsymbol{\Delta}_P = \boldsymbol{M}_{W_P}\boldsymbol{\Sigma}\boldsymbol{M}_{W_P}'$ to reflect the distinct weighting structures of the benchmark- and portfolio-weighted regressions. In practice, however, using a single residual covariance $\boldsymbol{\Delta}$, typically based on $\boldsymbol{W}_B$ for consistency with benchmark-relative reporting, yields nearly identical results for benchmark-aware portfolios.

Process summary

In summary, we compute complete Brinson-style attribution with zero interaction terms as follows.

1. Run the two unrestricted WLS return regressions under benchmark and portfolio weights.
2. Obtain the return covariance $\widehat{\boldsymbol{\Sigma}}$ and construct the residual covariance $\widehat{\boldsymbol{\Delta}} = \boldsymbol{M}_W \widehat{\boldsymbol{\Sigma}} \boldsymbol{M}_W'$.
3. Propagate $\widehat{\boldsymbol{\Delta}}$ to the group-mean covariances $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_P$, and $\boldsymbol{\Sigma}_{B,P}$, and assemble the joint covariance $\boldsymbol{\Gamma}$.
4. Apply the restricted GLS update (20) to obtain the adjusted group means $(\widehat{\boldsymbol{\mu}}_B^*, \widehat{\boldsymbol{\mu}}_P^*)$ that satisfy $\boldsymbol{w}_A' \boldsymbol{D}(\widehat{\boldsymbol{\mu}}_P^* - \widehat{\boldsymbol{\mu}}_B^*) = 0$.
5. Compute the adjusted attribution components $A^* = \boldsymbol{w}_A'(\widehat{\boldsymbol{r}}_B^* - r_B \boldsymbol{\iota})$ and $S^* = \boldsymbol{w}_B'(\widehat{\boldsymbol{r}}_P^* - \widehat{\boldsymbol{r}}_B^*) = S + I - (A^* - A)$, so that $r_A = A^* + S^*$ with no interaction term.

The procedure can be viewed as a two-step solution: first estimate the benchmark- and portfolio-weighted group means as in the standard Brinson decomposition, then adjust both by restricted GLS so that the active return identity holds exactly. The result preserves the familiar structure of Brinson attribution while reallocating the interaction term in a statistically optimal, least-distorting manner.

Next, we show that for similar portfolio and benchmark weights, the GLS weights approach equality and the interaction is reallocated about 50/50 across allocation and selection components. When the portfolio is substantially more concentrated (or more diversified) than the benchmark, the GLS weighting tilts toward the more precise regression, reallocating more of the interaction toward the allocation (or selection) component, respectively.

4 Illustrative Examples

Consider $G = 2$ groups with benchmark weights $w_{B,1} = w_{B,2} = 0.5$ and portfolio weights $w_{P,1} = 0.6$, $w_{P,2} = 0.4$, so $w_{A,1} = 0.1$ and $w_{A,2} = -0.1$. Let the (unrestricted) group means be

$$\mu_{B,1} = 2.0\%, \qquad \mu_{B,2} = -1.0\%, \qquad \mu_{P,1} = 2.4\%, \qquad \mu_{P,2} = -1.2\%.$$

Then

$$r_B = 0.5(2.0\%) + 0.5(-1.0\%) = 0.50\%,$$
$$r_P = 0.6(2.4\%) + 0.4(-1.2\%) = 0.96\%,$$

so $r_A = r_P - r_B = 0.46\%$. The classical Brinson components are

$$A = \sum_g (w_{P,g} - w_{B,g})(\mu_{B,g} - r_B) = 0.30\%,$$
$$S = \sum_g w_{B,g}(\mu_{P,g} - \mu_{B,g}) = 0.10\%,$$
$$I = \sum_g (w_{P,g} - w_{B,g})(\mu_{P,g} - \mu_{B,g}) = 0.06\%,$$

so $A + S + I = r_A$, as required.

4.1 Balanced precision

If the two regressions have balanced precision, $\boldsymbol{\Gamma}$ is approximately block-diagonal with equal blocks, $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma})$, and negligible cross-covariance. When both estimates are equally precise, the interaction effect should be split evenly between allocation and selection. We can verify that the restricted GLS update indeed produces this outcome.

In this case,

$$\boldsymbol{H}\,\boldsymbol{\Gamma}\,\boldsymbol{H}' = 2\boldsymbol{\Sigma}, \qquad \boldsymbol{\Gamma}\,\boldsymbol{H}' = \begin{bmatrix} -\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} \end{bmatrix}$$

with $\boldsymbol{H} = [-\boldsymbol{I}_G \ \boldsymbol{I}_G]$.

Write the interaction scalar I as the deviation from the constraint

$$I \equiv \boldsymbol{w}_A' \boldsymbol{D}(\widehat{\boldsymbol{\mu}}_P - \widehat{\boldsymbol{\mu}}_B) = \boldsymbol{w}_A' \boldsymbol{D}\,\boldsymbol{H}\,\widehat{\boldsymbol{m}}.$$

We can plug these into the update equation

$$\widehat{\boldsymbol{m}}^* = \widehat{\boldsymbol{m}} - \boldsymbol{\Gamma}\,\boldsymbol{H}'\boldsymbol{D}'\boldsymbol{w}_A \left(\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\,\boldsymbol{\Gamma}\,\boldsymbol{H}'\boldsymbol{D}'\boldsymbol{w}_A\right)^{-1} \boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\widehat{\boldsymbol{m}}$$

and notice that the central term $\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{H}\,\boldsymbol{\Gamma}\,\boldsymbol{H}'\boldsymbol{D}'\boldsymbol{w}_A = 2\,\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A$ is a scalar. This gives the block adjustments

$$\Delta\boldsymbol{\mu}_B \equiv \widehat{\boldsymbol{\mu}}_B^* - \widehat{\boldsymbol{\mu}}_B = \frac{\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A}{2\,\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A}\, I,$$
$$\Delta\boldsymbol{\mu}_P \equiv \widehat{\boldsymbol{\mu}}_P^* - \widehat{\boldsymbol{\mu}}_P = -\frac{\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A}{2\,\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A}\, I.$$

The term $\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A \left(2\,\boldsymbol{w}_A' \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A\right)^{-1}$ plays the same role as the regression coefficient of the group-mean vector on the fitted active return. The numerator $\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{w}_A$ is the covariance between the estimated group means and the active fitted return, and the denominator is twice its variance. Hence,

the update to each block can be viewed as a regression-style adjustment: the group means move in proportion to their beta with the active fitted return and in proportion to the distance from the active-exact constraint. Group-mean components that are estimated with higher uncertainty (larger variance in $\boldsymbol{\Gamma}$) or that covary more with the active return absorb a larger share of the adjustment, just as in a standard GLS update.

The updated allocation effect is $A^* = \boldsymbol{w}_A' \widehat{\boldsymbol{r}}_B^* = \boldsymbol{w}_A' \boldsymbol{D} \widehat{\boldsymbol{\mu}}_B^*$, and the update to the allocation component is

$$\Delta A = \boldsymbol{w}_A' \boldsymbol{D}\, \Delta \boldsymbol{\mu}_B = \frac{\boldsymbol{w}_A' \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{D}' \boldsymbol{w}_A}{2\, \boldsymbol{w}_A' \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{D}' \boldsymbol{w}_A}\, I = \tfrac{1}{2}\, I.$$

By exactness, after the update $A^* + S^* = A + S + I$, so

$$\Delta S = I - \Delta A = \tfrac{1}{2}\, I.$$

Therefore, under balanced precision the restricted GLS update splits the interaction approximately 50/50,

$$A^* = A + \tfrac{1}{2} I = 0.33\%, \quad S^* = S + \tfrac{1}{2} I = 0.13\%,$$

with $I^* = 0$ by construction. This produces complete attribution with $A^* + S^* = r_A = 0.46\%$.[4]

If the benchmark and portfolio mean estimates have unequal precision or non-negligible cross-covariance, the same algebra yields unequal shares that are determined by the blocks of $\boldsymbol{\Gamma}$ (see section 3.4).

4.2 Spherical specific risk and optimal interaction splits

For cases where the reallocation may not be equal, we now consider cases where regression precision is determined mainly by the dispersion of portfolio weights. This occurs when returns have equal residual risk but portfolios differ in concentration. In this setting, differences in diversification alone determine the relative precision of the benchmark- and portfolio-weighted regressions, which in turn determines the optimal reallocation of the interaction term. The spherical risk case extends the balanced-precision logic by allowing the benchmark and portfolio regressions to differ in effective sampling precision because of unequal diversification.

[4] If the portfolio and benchmark employ different leverage, $\boldsymbol{w}_A' \boldsymbol{\iota} \neq 0$, the updated allocation effect includes a funding/leverage adjustment, $\Delta A = \tfrac{1}{2} I - (\boldsymbol{w}_A' \boldsymbol{\iota})\, \Delta r_B$, where $\Delta r_B = (\boldsymbol{w}_B' \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{D}' \boldsymbol{w}_A) / (2\, \boldsymbol{w}_A' \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{D}' \boldsymbol{w}_A)\; I$. Many reports show active attribution on a leverage-matched basis, in which case this term is zero by construction.

Under spherical residual risk, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_N$, each group-level regression has sampling variance that depends only on the dispersion of its weights. The benchmark- and portfolio-weighted group mean estimates therefore have block-diagonal covariance

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Sigma}_B & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_P \end{bmatrix},$$

with $\boldsymbol{\Sigma}_B = \sigma^2 \operatorname{diag}(s_{B,1}, \ldots, s_{B,G})$ and $\boldsymbol{\Sigma}_P = \sigma^2 \operatorname{diag}(s_{P,1}, \ldots, s_{P,G})$, where $s_{B,g} \propto \sum_{i \in g} w_{B,i}^2$ and $s_{P,g} \propto \sum_{i \in g} w_{P,i}^2$ measure the effective concentration of the benchmark and portfolio within group g.

Starting from the restricted GLS update,

$$\widehat{\boldsymbol{m}}^* = \widehat{\boldsymbol{m}} - \boldsymbol{\Gamma H}' \boldsymbol{D}' \boldsymbol{w}_A \left(\boldsymbol{w}_A' \boldsymbol{D H \Gamma H}' \boldsymbol{D}' \boldsymbol{w}_A \right)^{-1} \boldsymbol{w}_A' \boldsymbol{D H} \widehat{\boldsymbol{m}},$$

with $\boldsymbol{H} = [-\boldsymbol{I}_G \ \boldsymbol{I}_G]$ and $\widehat{\boldsymbol{m}}' = [\widehat{\boldsymbol{\mu}}_B', \widehat{\boldsymbol{\mu}}_P']$, as before. Define the scalar interaction as the deviation from the restriction, $I = \boldsymbol{w}_A' \boldsymbol{D} (\widehat{\boldsymbol{\mu}}_P - \widehat{\boldsymbol{\mu}}_B)$. Using $\boldsymbol{\Gamma H}' = [-\boldsymbol{\Sigma}_B; \boldsymbol{\Sigma}_P]$ and $\boldsymbol{H \Gamma H}' = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_P$, we obtain the block updates

$$\Delta \boldsymbol{\mu}_B = \quad \frac{\boldsymbol{\Sigma}_B \, D' \boldsymbol{w}_A}{\boldsymbol{w}_A' D (\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_P) D' \boldsymbol{w}_A} \, I,$$

$$\Delta \boldsymbol{\mu}_P = -\frac{\boldsymbol{\Sigma}_P \, D' \boldsymbol{w}_A}{\boldsymbol{w}_A' D (\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_P) D' \boldsymbol{w}_A} \, I.$$

The corresponding adjustments to the allocation and selection effects are

$$A^* = A + \lambda_A I, \qquad S^* = S + \lambda_S I, \qquad \lambda_A + \lambda_S = 1,$$

where the optimal reallocation weights follow directly from from the formulas above as

$$\lambda_A = \frac{\boldsymbol{w}_A' D \boldsymbol{\Sigma}_B D' \boldsymbol{w}_A}{\boldsymbol{w}_A' D (\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_P) D' \boldsymbol{w}_A}, \qquad \lambda_S = \frac{\boldsymbol{w}_A' D \boldsymbol{\Sigma}_P D' \boldsymbol{w}_A}{\boldsymbol{w}_A' D (\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_P) D' \boldsymbol{w}_A}.$$

If the portfolio has uniform relative concentration $s_{P,g} = k \, s_{B,g}$ across groups, the ratio

$$k \equiv \frac{\boldsymbol{w}_P' \boldsymbol{w}_P}{\boldsymbol{w}_B' \boldsymbol{w}_B}$$

fully characterizes the relative sampling precision of the two regressions.[5]

[5] The Brinson regressions must use the portfolio's actual (fully levered) weights to reproduce the correct group-wise means. The relative concentration index k appropriately reflects leverage:

The weight expressions simplify to

$$\lambda_S = \frac{1}{1+k}, \qquad \lambda_A = \frac{k}{1+k}.$$

Under spherical return risk, the concentration ratio k measures the relative sampling variance of the portfolio- and benchmark-weighted regressions. When $k > 1$, the portfolio is more concentrated and its group means are estimated less precisely; when $k < 1$, the portfolio is more diversified and its group means are estimated more precisely. Because the selection effect depends on the difference between the two regressions, while the allocation effect depends only on the benchmark, changes in portfolio concentration affect the noise in selection more strongly than in allocation. As concentration rises, the GLS update assigns a larger share of the interaction to allocation; as concentration falls, it assigns a larger share to selection.

Figure 1 shows this relationship for several concentration levels. When $k = 1$ (equal diversification), the split is 50/50. Even moderate concentration tilts the optimal redistribution substantially: if the portfolio holds only half the benchmark's names per sector, the interaction shifts from a 50/50 split to roughly one-third selection and two-thirds allocation. This statistical asymmetry grows with relative portfolio concentration, reflecting that more concentrated portfolios provide less precise within-group evidence and therefore merit smaller selection adjustments. However, it requires extreme differences in portfolio concentration to make a 100% reallocation of the interaction effect to either allocation or selection statistically optimal.

To illustrate numerically, suppose the portfolio is approximately twice as concentrated as the benchmark ($k = 2$). The GLS weights are then $\lambda_S = 1/3$ and $\lambda_A = 2/3$, implying

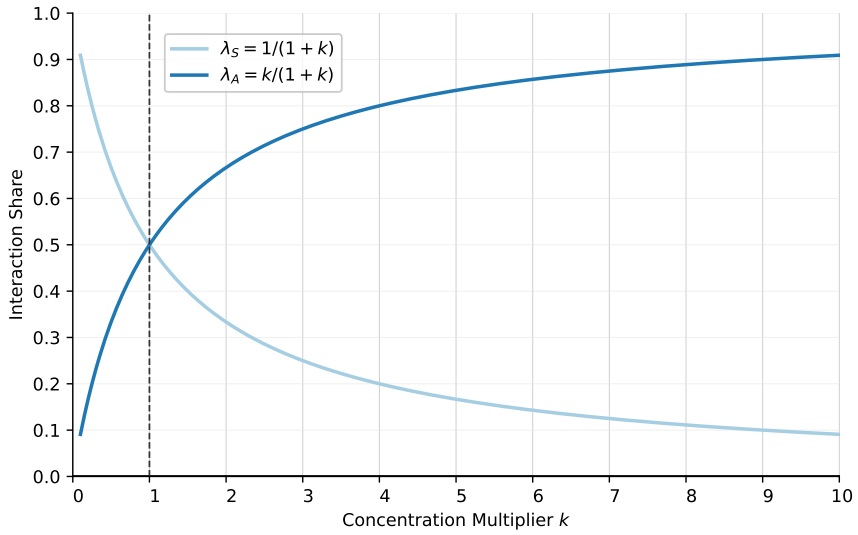
$$A^* = A + \tfrac{2}{3}I = 0.34\%,$$
$$S^* = S + \tfrac{1}{3}I = 0.12\%.$$

The adjusted effects again sum to $r_A = 0.46\%$, but the reallocation now places greater emphasis on the allocation component. Such asymmetry arises automatically from the relative statistical precision of the two regressions, without any arbitrary manual adjustment. The same logic applies symmetrically when the portfolio is more diversified than the benchmark, in which case the selection component absorbs a larger share of the interaction.

---

higher gross exposure increases the sampling variance of the portfolio-weighted regression and thus tilts the optimal reallocation.

Figure 1: Optimal GLS Split of Interaction Effects



The figure shows the statistically optimal distribution of the interaction effect I across allocation A and selection S. The optimal split adds a share λ_A of the interaction effect to the allocation component and λ_S to the selection component, so that the updated allocation and selection effects are

$$A^* = A + \lambda_A I \qquad \text{and} \qquad S^* = S + \lambda_S I.$$

If returns have spherical risk, $\boldsymbol{\Delta} = \sigma^2 I$, the concentration multiplier k measures relative portfolio diversification and is defined as

$$k = (\boldsymbol{w}_P' \boldsymbol{w}_P)/(\boldsymbol{w}_B' \boldsymbol{w}_B),$$

where $\boldsymbol{w}_P$ and $\boldsymbol{w}_B$ are the portfolio and benchmark weights, respectively. Lower k values correspond to more diversified portfolios, while higher values correspond to more concentrated portfolios.

The optimal reallocation shares are

$$\lambda_A = k/(k+1) \qquad \text{and} \qquad \lambda_S = 1/(k+1),$$

so that we distribute more of the interaction term to allocation for more concentrated portfolios, which contain more estimation error.

When concentration ratios differ across groups, the restricted GLS update reallocates the interaction term heterogeneously across them. Now, each group g has its own effective precision ratio

$$k_g = \frac{s_{P,g}}{s_{B,g}} = \frac{\sum_{i \in g} w_{P,i}^2}{\sum_{i \in g} w_{B,i}^2},$$

so that

$$\lambda_{S,g} = \frac{1}{1+k_g}, \qquad \lambda_{A,g} = \frac{k_g}{1+k_g}.$$

Groups with more concentrated portfolio weights ($k_g > 1$) receive a larger reallocation toward allocation and a smaller one toward selection, while more diversified groups ($k_g < 1$) receive the opposite treatment. If we define $\boldsymbol{a}_g = D'\boldsymbol{w}_{A,g}$, the overall portfolio-level split is

$$\lambda_S = \frac{\sum_g \boldsymbol{a}_g' \boldsymbol{\Sigma}_{P,g} \boldsymbol{a}_g}{\sum_g \boldsymbol{a}_g' (\boldsymbol{\Sigma}_{B,g} + \boldsymbol{\Sigma}_{P,g}) \boldsymbol{a}_g}, \qquad \lambda_A = 1 - \lambda_S,$$

which is a weighted average of these group-level shares, determined by each group's contribution to active risk. This formulation highlights that the GLS adjustment acts locally within groups, with each group's share of the interaction depending on its own relative precision. Note, however, that the residual interaction effects generally remain nonzero within each group. Elimination of interaction terms for all groups requires active-exactness constraints for each group rather than the single portfolio-wide restriction we use here. Appendix A discuss the group-specific constraints.

5 Summary

Both factor attribution and Brinson attribution can be expressed as regression problems. We can redistributed the nuisance residuals or interaction terms by applying a linear restriction to the regressions. This ensures that the attribution is complete and economically interpretable. The restricted regression framework provides a unified and flexible foundation for complete portfolio return attribution.

The statistical results suggest that, for portfolios that remain reasonably close to their benchmarks, splitting the interaction term equally between allocation and selection effects is typically near optimal. When portfolios are materially more concentrated than their benchmarks, holding fewer or larger positions within the same universe, the reallocation tends to tilt toward the allocation component. This reflects greater uncertainty in the within-group (selection) estimates. Conversely, when portfolios include securities outside the benchmark or are more diversified across names or markets, the reallocation may lean toward the selection component.

For multi-period attribution, one can aggregate period-by-period results or apply smoothing techniques such as those of Cariño (1999), Menchero (2000), and Frongello (2002), which adjust period returns so that the cumulative result matches the portfolio's multi-period performance. Naturally, the approach developed here suggests that cross-sectional and time-series restrictions applied simultaneously to panel regressions are an interesting alternative to these ad hoc smoothing techniques.

Panel regression methods can provide a complementary extension by imposing both cross-sectional and time-series restrictions. This enables arithmetic attribution that matches compounded returns through multi-period smoothing. Such time-series restrictions, closely related to the smoothing methods of Cariño (1999), Menchero (2000), and Frongello (2002), represent a natural direction for future work.

6 References

Bacon, Carl R., 2004, *Practical Portfolio Performance Measurement and Attribution* (John Wiley & Sons, Chichester, England).

Brinson, Gary P., and Nimrod Fachler, 1985, Measuring non-US equity portfolio performance, *Journal of Portfolio Management* 11 (3), 73–76.

Brinson, Gary P., L. Randolph Hood, and Gilbert L. Beebower, 1986, Determinants of portfolio performance, *Financial Analysts Journal* 42 (4), 39–44.

Cariño, David R., 1999, Combining attribution effects over time, *Journal of Performance Measurement* (Summer), 5–14.

Connor, Gregory, 1995, The three types of factor models: A comparison of their explanatory power, *Financial Analysts Journal* 51 (3), 42–46.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33 (1), 3–56.

Frongello, Andrew S.B., 2002, Linking single period attribution results, *Journal of Performance Measurement* (Spring), 10–22.

Greene, William H., and Terry G. Seaks, 1991, The restricted least squares estimator: A pedagogical note, *Review of Economics and Statistics* 73 (3), 563–567.

Grinold, Richard C., 2006, Attribution, *Journal of Portfolio Management* (Winter), 9–22.

Ledoit, Olivier, and Michael Wolf, 2004, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88 (2), 365–411.

Menchero, Jose G., 2000, An optimized approach to linking attribution effects over time, *Journal of Performance Measurement* (Fall), 36–42.

Sneddon, Leigh, 2021, Hidden errors in regression-based attribution, *Journal of Performance Measurement* (Winter), 10–17.

Spaulding, David, 2003, *Investment Performance Attribution* (McGraw–Hill, New York, NY).

Spaulding, David, 2008, Should the interaction effect be allocated? A "black box" approach to interaction, *Journal of Performance Measurement* (Spring), 23–30.

Theil, Henri, 1971, *Principles of Econometrics* (John Wiley & Sons, New York, NY).

A Group-Level Active-Exactness Restrictions

In the main text we imposed a single *active-exactness* restriction that jointly adjusts the benchmark- and portfolio-weighted regressions so that the portfolio-weighted active residuals vanish:

$$\boldsymbol{w}_A' \boldsymbol{D} \left(\widehat{\boldsymbol{\mu}}_P^* - \widehat{\boldsymbol{\mu}}_B^* \right) = 0. \tag{27}$$

This ensures that the portfolio's active return equals the sum of allocation and selection effects, with no residual or interaction term. In some reporting frameworks, investors also display allocation, selection, and interaction effects by sector. This appendix generalizes the single restriction to a set of group-level restrictions that eliminate interaction terms within each group.

A.1 Multiple active-exactness constraints

Let $g = 1, \ldots, G$ index groups (e.g., sectors), and define the diagonal selector $\boldsymbol{S}_g$ that isolates assets belonging to group g. Each $\boldsymbol{S}_g$ is an $N \times N$ diagonal matrix with ones for assets in group g and zeros elsewhere, so that $\boldsymbol{S}_g \boldsymbol{r}$ retains only the returns of assets in g. This choice preserves consistent vector dimensions across all groups.

We continue to define $\boldsymbol{w}_A = \boldsymbol{w}_P - \boldsymbol{w}_B$ as the vector of active portfolio weights and

$$\widehat{\boldsymbol{m}} \equiv \begin{bmatrix} \widehat{\boldsymbol{\mu}}_B \\ \widehat{\boldsymbol{\mu}}_P \end{bmatrix}, \qquad \boldsymbol{H} \equiv \begin{bmatrix} -\boldsymbol{I}_G & \boldsymbol{I}_G \end{bmatrix}, \tag{28}$$

so that $\boldsymbol{H}\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{\mu}}_P - \widehat{\boldsymbol{\mu}}_B$. For each group g, define $\boldsymbol{c}_g = \boldsymbol{S}_g \boldsymbol{w}_A$ as the active-weight vector restricted to assets in that group, and let $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_G]$ stack these vectors. The group-level active-exactness condition requires the portfolio-weighted residual in each group to vanish:

$$\boldsymbol{c}_g' \boldsymbol{D}\, \boldsymbol{H} \widehat{\boldsymbol{m}}^* = 0, \qquad g = 1, \ldots, G. \tag{29}$$

Stacking these restrictions yields

$$\boldsymbol{C}' \boldsymbol{D}\, \boldsymbol{H} \widehat{\boldsymbol{m}}^* = \boldsymbol{0}. \tag{30}$$

If a group has zero active weight ($w_{P,g} = w_{B,g}$), the corresponding column of $\boldsymbol{C}$ is zero and can be omitted. Summing all group constraints recovers the single portfolio-level restriction used in the main text.

A.2 Restricted GLS formulation

Let $\widehat{\boldsymbol{m}}$ denote the unrestricted stacked group-mean estimates obtained from the two separate regressions. To satisfy the G_c group-level active-exactness constraints $\boldsymbol{C}'\boldsymbol{D}\boldsymbol{H}\widehat{\boldsymbol{m}}^* = \boldsymbol{0}$ while minimally disturbing the unrestricted estimates, we solve the restricted GLS problem

$$\min_{\boldsymbol{m}} \tfrac{1}{2}(\boldsymbol{m} - \widehat{\boldsymbol{m}})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{m} - \widehat{\boldsymbol{m}}) \quad \text{s.t.} \quad \boldsymbol{C}'\boldsymbol{D}\boldsymbol{H}\boldsymbol{m} = \boldsymbol{0}, \tag{31}$$

where $\boldsymbol{\Gamma} = \mathrm{Var}(\widehat{\boldsymbol{m}})$ is the stacked covariance matrix of the unrestricted group-mean estimates, constructed exactly as in section 3.4 from the residual covariance $\boldsymbol{\Delta} = \boldsymbol{M}_W \boldsymbol{\Sigma} \boldsymbol{M}_W'$.

Once again, the standard restricted GLS solution is

$$\widehat{\boldsymbol{m}}^* = \widehat{\boldsymbol{m}} - \boldsymbol{\Gamma}\,(\boldsymbol{C}'\boldsymbol{D}\boldsymbol{H})'\left[\boldsymbol{C}'\boldsymbol{D}\boldsymbol{H}\,\boldsymbol{\Gamma}\,(\boldsymbol{C}'\boldsymbol{D}\boldsymbol{H})'\right]^{-1}\boldsymbol{C}'\boldsymbol{D}\boldsymbol{H}\,\widehat{\boldsymbol{m}}. \tag{32}$$

The updated stacked estimates $\widehat{\boldsymbol{m}}^* = (\widehat{\boldsymbol{\mu}}_B^*, \widehat{\boldsymbol{\mu}}_P^*)'$ jointly satisfy

$$\boldsymbol{C}'\boldsymbol{D}\,\boldsymbol{H}\widehat{\boldsymbol{m}}^* = \boldsymbol{0}, \tag{33}$$

so that every constrained group has zero interaction term. Because the portfolio total is the sum of all groups, the overall portfolio-level interaction term also vanishes.

A.3 Interpretation and practical remarks

This formulation generalizes the single portfolio-level active-exactness restriction in a natural way. The one-constraint system of the main text is recovered when $\boldsymbol{C} = \boldsymbol{c}$. Adding further constraints enforces zero interaction effects within each group while preserving the total active return $r_A = A + S$. Although these additional restrictions introduce slightly more distortion relative to the unrestricted estimates $\widehat{\boldsymbol{m}}$, the impact is typically small when $G \ll N$ and most groups contain many securities. The system (32) or its equivalent KKT form is symmetric indefinite and can be solved efficiently via an LDL' factorization, as recommended by Greene and Seaks (1991). This generalization is therefore both conceptually straightforward and practically useful for attribution systems that report sector-level effects.

B Long-Short and Market-Neutral Portfolios

Factor-based attribution extends naturally to longâĂŞshort and market-neutral portfolios. Negative portfolio or benchmark weights cause no

difficulties in computing factor exposures or attributing returns. For market-neutral portfolios, we can use notional allocations to define portfolio weights and proceed exactly as before. Once the portfolio weights are well defined, factor-based return attribution, whether unconstrained or under the linear restriction introduced in the main text, remains valid. Using a cash benchmark for a market-neutral portfolio is economically sensible and poses no technical problems for factor-based attribution.

Brinson-style attribution, however, requires additional care when portfolio allocations are zero or when the portfolio is market-neutral. In this case, the natural cash benchmark has zero exposure to all risky assets and produces zero returns in each group, so allocation and selection effects for risky assets vanish by construction. To recover economically meaningful attributions, many practitioners split market-neutral portfolios into separate long and short partitions. Standard attribution is then performed on the long side, and again on the absolute value of the short positions, both relative to the same long-only benchmark; the results are then combined (long minus short). In this formulation, a long-only benchmark provides a natural reference for both sides.

This split-book approach can also be used within the restricted regression framework developed in the main text. Each sleeve (long and short) can be analyzed separately using the same active-exactness restriction, and the resulting effects recombined to yield a coherent, additive attribution for the overall market-neutral portfolio.