The Best of Both: Hybrid Fundamental and Statistical Risk Models

Ludger Hentschel

August 31, 2025

Abstract

We extend a fundamental factor model by applying a statistical model to its residuals. This hybrid approach keeps the familiar, intuitive structure of the fundamental model while adding statistical factors to capture additional structure in the residual covariance. By ensuring that the statistical factors are orthogonal to the fundamental ones, we preserve the interpretability of the fundamental factors and at the same time improve estimates of residual risk.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

Acknowledgements
For stimulating this work and for helpful suggestions, I am grateful to Simon Bell, Deepak Gurnani, Nishant Gurnani, Chandra Prakash, Ernst Schaumburg, Sahil Singal, and Eric Vogt.

1 Introduction

We often use fundamental factor models because they are intuitive and tied directly to observable characteristics such as industries, styles, or exposures. These models provide a clear link between financial characteristics and the associated portfolio risk. In practice, the residual risk from a fundamental model may disguise patterns of common variation across assets. If these correlations exist but are not modeled, the model is likely to underestimate portfolio risk.

A hybrid model addresses this by applying a statistical risk model to the residuals from a fundamental factor model. The statistical factors capture structure in the residual covariance that the fundamental model leaves in the idiosyncratic term. By orthogonalizing these statistical factors with respect to the fundamental exposures, we leave the fundamental factors, their returns, and their covariance completely unchanged. This preserves the interpretability of the fundamental model while improving the accuracy of risk estimates.

Such a hybrid model provides risk estimates and risk attribution just like a purely fundamental model, except that a few of the factors are statistical and are not directly associated with observable characteristics.[1]

A purely fundamental model also enables intuitive return attribution. In the hybrid model, we continue to use the fundamental factors for return attribution. This fundamental return attribution is not affected by the addition of the statistical components since the fundamental components remain unchanged by design. Unfortunately, the statistical factors have arbitrary signs and can rotate period by period. This unavoidable ambiguity makes multi-period return attribution with statistical factors difficult to interpret. The statistical components are best understood as tools for risk estimation, not for interpreting realized returns.

By constructing the statistical risk model for the residuals from the fundamental factor model we clearly give priority to the fundamental factor model. The fundamental factors are easier to interpret and easier to estimate. The only reason to add the statistical model is that we have not yet identified additional fundamental factors to capture the remaining structure. In this sense, the statistical factors are a convenient interim solution while we search for additional fundamental factors.

[1] The term hybrid factor model has been used in the literature in different ways. Menchero and Mitra (2008) combine fundamental definitions for some factors with statistical estimation for others. The approach here differs: we retain a full set of fundamental factors and then apply a statistical model to the residuals. This may be similar to the risk model structure of the "Everything Everywhere" risk model by Northfield Information Services; but details of this model are not publicly available.

The next sections summarize fundamental and statistical risk models, respectively. Section 4 describes in detail how to combine these two approaches. Section 5 concludes.

2 Fundamental Factor Model

Following Connor (1995), we construct a fundamental factor model for N assets. Let $\boldsymbol{r}_{t+1}$ denote the $(N \times 1)$ vector of asset returns from time t to $t+1$. Assume that the returns follow a K-factor fundamental model of the form

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

where $\boldsymbol{X}_t$ is the $(N \times K)$ matrix of known factor exposures at time 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 factors $\boldsymbol{X}_t$ can include a wide variety of security characteristics. We can include market exposures, either as dummy variables or as betas. In similar form, we can include industry or asset class exposures. We can include security characteristics like company size, valuation ratios, past return patterns like momentum or reversals, or bond duration. We generally choose these characteristics because we can demonstrate that they contribute to risk shared across the assets. We can also include characteristics that we use to predict returns and construct portfolios, since these generally also contribute to risk. Examples of this include proprietary valuation measures.

2.1 Factor Risk Model

Conventionally, fundamental factor models assume that the residual returns $\boldsymbol{\varepsilon}_{t+1}$ are mean-zero, uncorrelated across assets, and uncorrelated with the fundamental factor returns $\boldsymbol{b}_{t+1}$. Under these assumptions, the return covariance matrix is

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

where $\boldsymbol{\Omega}_t$ is the $(K \times K)$ covariance matrix of factor returns and $\boldsymbol{D}_t$ is the $(N \times N)$ diagonal matrix of asset-specific residual variances. In this decomposition, all systematic co-movement is captured by the factor structure $\boldsymbol{X}_t \boldsymbol{\Omega}_t \boldsymbol{X}_t'$, while $\boldsymbol{D}_t$ models risk that is unique to each asset.

With the covariance matrix $\boldsymbol{\Sigma}_t$, we can estimate and manage total risk for a portfolio, estimate and manage portfolio risk contributions from the different factors, and use these risk estimates in portfolio construction.

2.2 Pure Factor Portfolios

From the exposures $\boldsymbol{X}_t$, we can construct K pure factor portfolios with unit exposure to each factor and zero exposure to all others. One approach is

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

which yields the ordinary least squares (OLS) factor mimicking portfolios. A more general, and often preferred, approach applies weighted least squares (WLS) to find

$$\boldsymbol{W}_t = \boldsymbol{\Gamma}_t\boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{X}_t)^{-1}, \tag{4}$$

where $\boldsymbol{\Gamma}_t$ is a diagonal matrix of asset weights, such as inverse residual variances $\boldsymbol{\Gamma}_t = \boldsymbol{D}_t^{-1}$ or market capitalization weights.

The factor $(\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{X}_t)^{-1}$ standardizes and decorrelates the exposures $\boldsymbol{X}_t$, so that column j of $\boldsymbol{W}_t$ is the $(N \times 1)$ factor mimicking portfolio with unit exposure to factor j and zero exposure to all other factors,

$$\boldsymbol{W}_t'\boldsymbol{X}_t = \boldsymbol{I}_K. \tag{5}$$

Any portfolio with exposures in the column space of $\boldsymbol{X}_t$ can be written as a linear combination of these pure factor portfolios.

Because $\boldsymbol{X}_t$ is observable at time t, these portfolios are known in advance and can be held to realize the corresponding factor return for the next period

$$\begin{aligned}
\widehat{\boldsymbol{b}}_{t+1} &= \boldsymbol{W}_t'\boldsymbol{r}_{t+1} && (6)\\
&= (\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{r}_{t+1} && (7)\\
&= \boldsymbol{b}_{t+1} + (\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{\Gamma}_t\boldsymbol{\varepsilon}_{t+1}. && (8)
\end{aligned}$$

Of course, neither the security returns $\boldsymbol{r}_{t+1}$ nor the factor returns $\boldsymbol{b}_{t+1}$ are known in advance.

The return we earn is equal to the true factor return $\boldsymbol{b}_{t+1}$ plus some noise that we were unable to diversify away. The main reasons for weighting the data are to estimate the true factor returns more precisely and to earn the true factor return with less risk.

2.3 Covariance Estimates

In order to populate the covariance matrix in equation (2), we need estimates of the factor covariance $\boldsymbol{\Omega}_t$ and the idiosyncratic variances $\boldsymbol{D}_t$. We already have the factor exposures $\boldsymbol{X}_t$.

By inspection of equation (7), we can see that the factor return estimates are equal to regression estimates of $\boldsymbol{b}_{t+1}$ from a multivariate regression of $\boldsymbol{r}_{t+1}$ on $\boldsymbol{X}_t$, using weights $\boldsymbol{\Gamma}_t^{1/2}$ for all observations. These regressions are familiar from Fama and MacBeth (1973).

The variances in $\boldsymbol{D}_t$ are the variances of the regression residuals

$$\widehat{\boldsymbol{\varepsilon}}_t = \boldsymbol{r}_{t+1} - \boldsymbol{X}_t \widehat{\boldsymbol{b}}_t \tag{9}$$

$$= (\boldsymbol{I}_N - \boldsymbol{X}_t \boldsymbol{W}_t')\boldsymbol{r}_{t+1}. \tag{10}$$

For variance or risk estimates, we commonly use exponentially weighted estimates.[2] This approach usefully approximates the ARCH and GARCH structure of security variances documented by Engle (1982) and Bollerslev (1986). We can apply these weighted averages to estimate the sample covariance of $\widehat{\boldsymbol{b}}_t$ and the sample variances of $\widehat{\boldsymbol{\varepsilon}}_t$.

2.4 Announcement Effects

If we wish to model event-time risk associated with announcements of earnings or macroeconomic data, it is best to make these adjustments at the level of the fundamental model. We may include an earnings announcement factor or adjust the time-series model of residual volatility. Accounting for this structure before applying a statistical risk model improves the stationarity of the residuals and helps the statistical model extract persistent structure. The statistical risk model itself cannot anticipate earnings announcement dates.

3 Statistical Risk Model

Following Connor and Korajczyk (1986) and Connor (1995), a statistical risk model provides a low-rank approximation to the covariance matrix by extracting a small number of principal components (PCs) from returns. Once again, the idea is to summarize co-movements in asset returns with a few factors, leaving only asset-specific risk in the residuals. Here, however, the common factors are latent and not specified explicitly. Instead, we estimate the factor exposures and factor returns.

If weighting of the observed returns is helpful in estimating the fundamental factor model, it is helpful for any estimation using the returns. The primary objective is to reduce heteroskedasticity across observations. This is accomplished with the same weights. As a result, we apply the same observation-specific weights here. Let $\boldsymbol{\Gamma}_t$ be the positive definite $(N \times N)$

[2] For example, see the RiskMetrics description in J.P. Morgan/Reuters (1996).

weighting matrix at time t, and define the whitened returns

$$\underline{\boldsymbol{r}}_{t+1} = \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{r}_{t+1}. \tag{11}$$

We use the underline notation to indicate that $\underline{\boldsymbol{r}}_{t+1}$ is the whitened version of the corresponding variable $\boldsymbol{r}_{t+1}$ and underline all variables associated with the whitened return space.

In practice, $\boldsymbol{\Gamma}_t$ is usually diagonal, with entries chosen to standardize residual variances from the fundamental regression (for example, setting $\Gamma_{ii,t}$ equal to the inverse of the residual variance for asset i). This whitening step does not change the span of the residuals but rescales them so that each contributes comparably. By doing so, the statistical model emphasizes structure in the correlations rather than being dominated by a few assets with high residual volatility.

Stacking a T-period sample of such vectors gives the $(N \times T)$ matrix of weighted returns $\underline{\boldsymbol{R}}$. The cross-sectionally weighted sample covariance is then

$$\widehat{\underline{\boldsymbol{\Sigma}}}_t = \frac{1}{T} \underline{\boldsymbol{R}}\, \underline{\boldsymbol{R}}', \tag{12}$$

which replaces the unweighted covariance $\boldsymbol{R}\boldsymbol{R}'/T$. As before, we can apply exponential weighting to the covariance estimate.

3.1 Principal Components

We now decompose $\widehat{\underline{\boldsymbol{\Sigma}}}_t$ into its principal components. Let $\underline{\boldsymbol{Z}}_t$ denote the $(N \times k)$ matrix of the top k eigenvectors, and let

$$\underline{\boldsymbol{\Lambda}}_t = \operatorname{diag}(\underline{\lambda}_{1t}, \ldots, \underline{\lambda}_{kt}) \tag{13}$$

collect the associated eigenvalues. By construction, the eigenvectors are orthonormal, so that

$$\underline{\boldsymbol{Z}}_t' \underline{\boldsymbol{Z}}_t = \boldsymbol{I}_k. \tag{14}$$

As for the fundamental factor model, we choose a number of factors k that is materially smaller than the number of assets N.

Under this model, we approximate the covariance of whitened returns using k statistical factors as

$$\widehat{\underline{\boldsymbol{\Sigma}}}_t = \underline{\boldsymbol{Z}}_t \underline{\boldsymbol{\Lambda}}_t \underline{\boldsymbol{Z}}_t' + \underline{\boldsymbol{\Delta}}_t, \tag{15}$$

where $\underline{\boldsymbol{\Delta}}_t$ is a diagonal matrix of residual variances.

The diagonal factor covariance $\underline{\boldsymbol{\Lambda}}_t$ is especially useful here because the signs of the eigenvectors $\underline{\boldsymbol{Z}}_t$ are indeterminate. A random change in sign has no effect on squared variance terms but makes the covariance estimates unreliable, unless they are zero.

When we use $k = N$ statistical factors, the factor model fully describes the covariance matrix and $\underline{\boldsymbol{\Delta}} = \mathbf{0}$. If we use fewer statistical factors, $k < N$, then allowing for idiosyncratic risk improves the approximation to the covariance.

The associated factor model for the whitened returns once again takes the form

$$\underline{\boldsymbol{r}}_{t+1} = \underline{\boldsymbol{Z}}_t \underline{\boldsymbol{f}}_{t+1} + \underline{\boldsymbol{e}}_{t+1}, \tag{16}$$

where $\underline{\boldsymbol{e}}_{t+1}$ are whitened residuals because they are derived from whitened returns. The projected returns $\underline{\boldsymbol{Z}}_t' \underline{\boldsymbol{r}}_{t+1}$ are not, by default, factor returns in the usual sense because $\underline{\boldsymbol{Z}}_t$ is often estimated using a sample that includes $\underline{\boldsymbol{r}}_{t+1}$. This makes investments conditional on $\underline{\boldsymbol{Z}}_t$ infeasible at time t. However, if $\underline{\boldsymbol{Z}}_t$ is based only on data through t, then

$$\widehat{\underline{\boldsymbol{f}}}_{t+1} = \underline{\boldsymbol{Z}}_t' \underline{\boldsymbol{r}}_{t+1} \tag{17}$$

is a factor return. Since $\underline{\boldsymbol{Z}}_t$ is orthonormal, the factor exposures are equal to the factor portfolio weights.

3.2 Mapping Back to Raw Returns

Since we base the entire estimation on whitened returns, the resulting covariance matrix is for the whitened returns. To move back to raw returns, we use

$$\boldsymbol{r}_{t+1} = \boldsymbol{\Gamma}_t^{-1/2} \underline{\boldsymbol{r}}_{t+1}, \tag{18}$$

with covariance

$$\begin{aligned} \boldsymbol{\Sigma}_t &= \boldsymbol{\Gamma}_t^{-1/2} \underline{\boldsymbol{\Sigma}}_t \boldsymbol{\Gamma}_t^{-1/2} && (19) \\ &= \boldsymbol{\Gamma}_t^{-1/2} \underline{\boldsymbol{Z}}_t \underline{\boldsymbol{\Lambda}}_t \underline{\boldsymbol{Z}}_t' \boldsymbol{\Gamma}_t^{-1/2} + \boldsymbol{\Gamma}_t^{-1/2} \underline{\boldsymbol{\Delta}}_t \boldsymbol{\Gamma}_t^{-1/2}. && (20) \end{aligned}$$

3.3 The Number of Components

In practice, the number of statistical factors k must be chosen carefully to balance model complexity with estimation error. Common heuristics include stopping at the "elbow" of the scree plot of eigenvalues, setting

a minimum variance threshold (e.g., explain 50% of total variance), using information criteria designed for factor models, or optimizing the cross-validated forecasts. The main requirement is that $k \ll N$.

When working with a sample containing T periods, the most common approach is to choose the smallest k such that the top k eigenvalues explain a target proportion of the total sample residual variance

$$\frac{\sum_{j=1}^{k} \underline{\lambda}_{j,t}}{\sum_{j=1}^{m} \underline{\lambda}_{j,t}} \geq \ell, \qquad \text{with} \quad m = \min\{N, T\}. \tag{21}$$

We typically choose ℓ between 30% and 90%, and $\underline{\lambda}_{j,t}$ are the eigenvalues.

Here, the dual PCA interpretation of Connor and Korajczyk (1993) and Jones (2001) can be helpful when $T < N$. The nonzero eigenvalues of the $(T \times T)$ matrix $\underline{\boldsymbol{R}}'\underline{\boldsymbol{R}}/T$ are the same as those of the $(N \times N)$ matrix $\underline{\boldsymbol{R}}\,\underline{\boldsymbol{R}}'/T$. This formulation avoids explicit estimation of the full $(N \times N)$ covariance when $N > T$ and provides direct access to the nonzero spectrum.[3]

When T is not too small, one can adapt information criteria such as those proposed by Bai and Ng (2002). These penalize model complexity and have asymptotic justification under high-dimensional assumptions. Although they require estimation of more than k eigenvalues, they can also be applied on $\underline{\boldsymbol{R}}'\underline{\boldsymbol{R}}$ to avoid working with the full $(N \times N)$ matrix.

Ultimately, the choice of k should balance statistical parsimony, stability over time, and marginal improvement in risk estimation.

3.4 Eigenvalue Shrinkage

Marchenko and Pastur (1967) and Johnstone (2001) show that the eigenvalues of $\widehat{\underline{\boldsymbol{\Sigma}}}_t$ contain systematic errors when N is large relative to T: the largest eigenvalues are biased upward, the smallest are biased downward, and variance is misallocated across components.[4] To stabilize risk estimates, we can shrink eigenvalues before using them in risk forecasts or portfolio construction. This also suggests that we should either choose k based on the shrunk eigenvalues or iterate over choices of k and shrunk estimates. The shrinkage applies to the eigenvalues only. It does not affect the eigenvectors.

The nonlinear shrinkage methods of Ledoit and Wolf (2015) and Ledoit and Wolf (2020) shrink the estimated eigenvalues $\underline{\lambda}_{j,t}$ to $\widehat{\underline{\lambda}}_{j,t}$ using the

[3] If we also want to recover the N-dimensional eigenvectors corresponding to these eigenvalues, we can obtain them by premultiplying the $(N \times T)$ residual matrix $\underline{\boldsymbol{R}}$ with the T-dimensional eigenvectors of $\underline{\boldsymbol{R}}'\underline{\boldsymbol{R}}/T$ and re-normalizing. In the present context, however, we only require the eigenvalue spectrum to choose k.

[4] Whether we compute eigenpairs via the $(N \times N)$ covariance or its $(T \times T)$ dual has no effect on the precision of the eigenvalue estimates, it is purely a computational choice. The statistical accuracy of these estimates is determined entirely by the ratio N/T, not by which representation we use.

empirical eigenvalue spectrum. These procedures can be applied to the leading part of the spectrum (via the $(T \times T)$ dual PCA when $T < N$), and typically improve out-of-sample risk forecasts relative to unshrunk eigenvalues or simple linear rules, especially when ν is large. Unfortunately, they are somewhat complex.

A simpler mechanism applies linear shrinkage to the estimated eigenvalues. Let $\overline{\underline{\lambda}}_t$ denote the average variance per asset

$$\overline{\underline{\lambda}}_t = \frac{1}{N} \operatorname{Tr}\left(\widehat{\underline{\boldsymbol{\Sigma}}}_t\right). \tag{22}$$

We can use this as a shrinkage target for the estimated eigenvalues.[5]

For the top k retained components, define the shrunk eigenvalues as

$$\widehat{\underline{\lambda}}_{j,t} = (1-\theta)\underline{\lambda}_{j,t} + \theta\overline{\underline{\lambda}}_t, \qquad j = 1, \ldots, k. \tag{23}$$

The shrinkage intensity $\theta \in [0,1]$ controls the trade-off: $\theta = 0$ is no shrinkage, $\theta = 1$ collapses all retained eigenvalues to the grand mean, and intermediate values blend the individual estimates with the grand mean.

A practical rule of thumb ties θ to the dimension ratio $\nu = N/\tau$,

$$\theta \approx \frac{\nu}{1+\nu}, \tag{24}$$

where τ is the effective number of sample periods. If we equally weight all observations, then $\tau = T$. If we exponentially weight observations with exponential coefficient ϕ, then the effective number of observations is

$$\tau \approx \frac{1+\phi}{1-\phi}\,\frac{1-\phi^T}{1+\phi^T}. \tag{25}$$

When we use slow exponential decay, $\phi \approx 1$, so that $\tau \approx T$. When T becomes very large, $\tau \approx (1+\phi)/(1-\phi)$. For additional details, see Hentschel (2024).

If $N \ll T$, then ν is small and θ is close to 0 and we apply light shrinkage. If $N \approx T$, then $\theta \approx 0.5$. If $N \gg T$, θ is close to 1 and we apply stronger shrinkage. This rule is transparent, easy to implement, and robust in practice.

After shrinking, the working covariance is

$$\widehat{\underline{\boldsymbol{\Sigma}}}_t = \underline{\boldsymbol{Z}}_t \widehat{\underline{\boldsymbol{\Lambda}}}_t \underline{\boldsymbol{Z}}_t' + \underline{\boldsymbol{\Delta}}_t. \tag{26}$$

[5] When $T < N$ and we compute eigenpairs via the $(T \times T)$ dual PCA (eigendecomposition of $\underline{\boldsymbol{R}}'\underline{\boldsymbol{R}}/T$), the correct target remains $\overline{\underline{\lambda}}_t = \operatorname{Tr}(\widehat{\underline{\boldsymbol{\Sigma}}}_t)/N$, i.e., total variance divided by the number of assets. This can be computed directly from the sample variances (the trace) without forming the $(N \times N)$ matrix.

where $\underline{\boldsymbol{Z}}_t$ is composed of the top k eigenvectors of the covariance matrix $\widehat{\underline{\boldsymbol{\Sigma}}}_t$. The covariance $\widehat{\underline{\boldsymbol{\Lambda}}}_t = \text{diag}(\widehat{\underline{\lambda}}_{1t}, \ldots, \widehat{\underline{\lambda}}_{kt})$ contains the corresponding k largest, shrunk eigenvalues. In practice, shrinking the largest few eigenvalues and treating the remainder as idiosyncratic risk balances stability with parsimony and avoids overfitting. Such shrinkage affects the factor covariance but not the residuals and their variances.

3.5 Covariance Estimates

In order to populate the covariance matrix in equation (26), we need estimates of the factor covariance $\widehat{\underline{\boldsymbol{\Lambda}}}_t$ and the idiosyncratic variances $\underline{\boldsymbol{\Delta}}_t$. We have the factor exposures $\underline{\boldsymbol{Z}}_t$ and weights $\boldsymbol{\Gamma}_t$ for each period.

We can estimate the sample covariance $\widehat{\underline{\boldsymbol{\Sigma}}}_t$ of the period-by-period weighted returns $\underline{\boldsymbol{r}}_t$, possibly with exponential weights. From this covariance, we can then extract the eigenvalues and eigenvectors. The k largest, shrunk eigenvalues populate $\widehat{\underline{\boldsymbol{\Lambda}}}_t$.

We can estimate the residual covariance from the whitened residual returns

$$\widehat{\underline{\boldsymbol{e}}}_{t+1} = \underline{\boldsymbol{r}}_{t+1} - \underline{\boldsymbol{Z}}_t \widehat{\underline{\boldsymbol{f}}}_{t+1} \tag{27}$$

$$= \left(\boldsymbol{I}_N - \underline{\boldsymbol{Z}}_t \underline{\boldsymbol{Z}}_t'\right) \underline{\boldsymbol{r}}_{t+1}, \tag{28}$$

with $\underline{\boldsymbol{\Delta}}_t$ placing $\text{Var}_t(\widehat{\underline{e}}_{it+1})$ on the diagonal. Here too, we can apply exponential weights, if we wish.

After assembling the covariance for the whitened returns $\underline{\boldsymbol{r}}_t$, we can translate this back to the raw return space, according to equation (26).

4 Hybrid Risk Model

The hybrid model combines a fundamental factor model with a statistical factor model applied to the residuals. The guiding principle is to preserve the interpretation of fundamental factors while statistically capturing additional structure in the residual covariance. To achieve this, we orthogonalize the statistical exposures with respect to the fundamental exposures, which leaves the fundamental factors unchanged compared to a purely fundamental model.

For analytical and computational convenience, we construct the statistical model in the whitened space defined by the same weights $\boldsymbol{\Gamma}_t$ used in the fundamental model, and only map back to raw returns at the end.

4.1 Fundamental Residuals and Whitening

Let $\boldsymbol{X}_t$ be the $(N \times K)$ matrix of fundamental exposures, and define the whitened regressors $\underline{\boldsymbol{X}}_t = \boldsymbol{\Gamma}_t^{1/2}\boldsymbol{X}_t$ and whitened returns $\underline{\boldsymbol{r}}_{t+1} = \boldsymbol{\Gamma}_t^{1/2}\boldsymbol{r}_{t+1}$. The (weighted/GLS) projection onto the span of $\underline{\boldsymbol{X}}_t$ is

$$\underline{\boldsymbol{P}}_{X,t} = \underline{\boldsymbol{X}}_t\left(\underline{\boldsymbol{X}}_t'\underline{\boldsymbol{X}}_t\right)^{-1}\underline{\boldsymbol{X}}_t'. \tag{29}$$

The whitened residuals from the fundamental model are then

$$\underline{\boldsymbol{\varepsilon}}_{t+1} = \left(\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t}\right)\underline{\boldsymbol{r}}_{t+1}, \tag{30}$$

i.e., the component of $\underline{\boldsymbol{r}}_{t+1}$ orthogonal to the span of $\underline{\boldsymbol{X}}_t$.

We now apply a statistical factor model to these whitened residuals. The initial steps are identical to the discussion in the previous section.

4.2 Covariance of Whitened Residuals

Stacking T periods into the $(N \times T)$ matrix $\widehat{\underline{\boldsymbol{E}}} = [\widehat{\underline{\varepsilon}}_1, \ldots, \widehat{\underline{\varepsilon}}_T]$, we form the sample covariance

$$\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t} = \frac{1}{T}\widehat{\underline{\boldsymbol{E}}}\,\widehat{\underline{\boldsymbol{E}}}'. \tag{31}$$

This is the object on which we apply PCA to extract statistical components.

As before, we frequently use exponentially weighted averages to estimate empirical covariances. We can do this here. We denote the covariance estimate $\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}$, regardless of the estimation method.

4.3 PCA of the Covariance

We compute the eigenpairs of $\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}$ and then choose those associated with the k largest eigenvalues

$$\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}\check{\underline{\boldsymbol{z}}}_{j,t} = \underline{\lambda}_{j,t}\check{\underline{\boldsymbol{z}}}_{j,t}, \qquad j = 1, \ldots, k. \tag{32}$$

We can collect the eigenvectors $\check{\underline{\boldsymbol{Z}}}_t = (\check{\underline{\boldsymbol{z}}}_{1,t}, \ldots, \check{\underline{\boldsymbol{z}}}_{k,t})$, so that $\check{\underline{\boldsymbol{Z}}}_t'\check{\underline{\boldsymbol{Z}}}_t = \boldsymbol{I}_k$.

In order to choose k, we can inspect all N eigenvalues after shrinking them, as described in the previous section.

This gives us a statistical risk model for the whitened residuals from the fundamental factor model. The statistical factors explain residual returns that are uncorrelated with fundamental factor returns, on average. At a point in time, however, the statistical and fundamental factor exposures may be correlated with each other. To avoid double counting overlapping

exposures between the fundamental and statistical factors, we now orthogonalize the statistical factors with respect to the fundamental factors. This intentionally assigns priority to the fundamental factors because they are easier to interpret.

4.4 Orthogonality to Fundamental Factors

We project the statistical factors into the orthogonal complement of $\underline{\boldsymbol{X}}_t$

$$\underline{\boldsymbol{Z}}_t^{\perp} = \big(\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t}\big)\check{\underline{\boldsymbol{Z}}}_t. \tag{33}$$

By construction,

$$\underline{\boldsymbol{Z}}_t^{\perp\prime}\underline{\boldsymbol{X}}_t = \check{\underline{\boldsymbol{Z}}}_t'\big(\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t}\big)'\underline{\boldsymbol{X}}_t = \check{\underline{\boldsymbol{Z}}}_t'\boldsymbol{0} = \boldsymbol{0}. \tag{34}$$

Naturally, rotating the k statistical exposures so they are orthogonal to the K fundamental exposures requires degrees of freedom. We have the required degrees of freedom when $N > K + k$. This is true, unless we have an unusually small number of assets, N, or an exceptionally large number of fundamental factors K. In such cases, few degrees of freedom remain. This constrains the rotation but also makes it unlikely that the residuals contain material common structure.

Although these components are orthogonal to the fundamental factors, they are generally no longer orthogonal to each other. We correct this next.

4.5 Cleaning Up the Orthogonalized Exposures

We now transform these orthogonal components to produce statistical exposures that (i) are orthonormal, (ii) remain orthogonal to $\underline{\boldsymbol{X}}_t$, and (iii) have a diagonal $(k \times k)$ factor covariance, still in whitened space.

Step 1: Apply QR in the residual subspace.

A thin QR decomposition of $\underline{\boldsymbol{Z}}_t^{\perp}$ represents the matrix as the product of an orthonormal matrix $\widetilde{\underline{\boldsymbol{Z}}}_t$ and an upper triangular matrix $\underline{\boldsymbol{U}}_t$

$$\underline{\boldsymbol{Z}}_t^{\perp} = \widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{U}}_t, \qquad \widetilde{\underline{\boldsymbol{Z}}}_t'\widetilde{\underline{\boldsymbol{Z}}}_t = \boldsymbol{I}_k, \tag{35}$$

The matrix $\widetilde{\underline{\boldsymbol{Z}}}_t$ is $(N \times k)$ and $\underline{\boldsymbol{U}}_t$ is $(k \times k)$.[6] By design of the QR decomposition, $\widetilde{\underline{\boldsymbol{Z}}}_t$ is an orthonormal basis for the column space of $\underline{\boldsymbol{Z}}_t^{\perp}$.

[6] The thin QR decomposition finds smaller matrices, here $(N \times k)$ and $(k \times k)$, compared to the $(N \times N)$ and $(N \times k)$ matrices of the full QR decomposition.

The right multiplication by $\underline{\boldsymbol{U}}_t$ or its inverse says that $\underline{\boldsymbol{Z}}_t^{\perp}$ and $\widetilde{\underline{\boldsymbol{Z}}}_t$ share the same column space, so that $\widetilde{\underline{\boldsymbol{Z}}}_t$ is also orthogonal to $\underline{\boldsymbol{X}}_t$,

$$\widetilde{\underline{\boldsymbol{Z}}}_t'\underline{\boldsymbol{X}}_t = \left(\underline{\boldsymbol{Z}}_t^{\perp}\underline{\boldsymbol{U}}_t^{-1}\right)'\underline{\boldsymbol{X}}_t = \underline{\boldsymbol{U}}_t^{-1'}\underline{\boldsymbol{Z}}_t^{\perp'}\underline{\boldsymbol{X}}_t = \mathbf{0}. \tag{36}$$

Step 2: Re-diagonalize the within-subspace covariance.

These revised factor exposures have a $(k \times k)$ factor covariance $\widetilde{\underline{\boldsymbol{Z}}}_t'\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}\widetilde{\underline{\boldsymbol{Z}}}_t$. We can diagonalize this covariance by finding the full eigendecomposition

$$\widetilde{\underline{\boldsymbol{Z}}}_t'\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}\widetilde{\underline{\boldsymbol{Z}}}_t = \underline{\boldsymbol{Q}}_t\underline{\boldsymbol{\Pi}}_t\underline{\boldsymbol{Q}}_t', \qquad \underline{\boldsymbol{Q}}_t'\underline{\boldsymbol{Q}}_t = \boldsymbol{I}_k, \quad \underline{\boldsymbol{\Pi}}_t = \operatorname{diag}(\underline{\pi}_{1t}, \ldots, \underline{\pi}_{kt}). \tag{37}$$

Here, $\underline{\boldsymbol{Q}}_t$ contains the eigenvectors and $\underline{\boldsymbol{\Pi}}_t$ contains the eigenvalues.

Finally, we can define the matching rotated loadings

$$\widehat{\underline{\boldsymbol{Z}}}_t = \widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t. \tag{38}$$

Right-multiplication by $\underline{\boldsymbol{Q}}_t$ stays within the same residual subspace, so weighted orthogonality to $\underline{\boldsymbol{X}}_t$ is preserved; $\underline{\boldsymbol{Q}}_t$ being orthogonal keeps the columns orthonormal.

With this, we have achieved our objective of finding statistical factor exposures $\widehat{\underline{\boldsymbol{Z}}}_t$ that have diagonal within-subspace covariance

$$\widehat{\underline{\boldsymbol{Z}}}_t'\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}\widehat{\underline{\boldsymbol{Z}}}_t = (\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t)'\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}(\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t) = \underline{\boldsymbol{Q}}_t'\widetilde{\underline{\boldsymbol{Z}}}_t'\widehat{\underline{\boldsymbol{\Sigma}}}_{\varepsilon,t}\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t = \underline{\boldsymbol{Q}}_t'\underline{\boldsymbol{Q}}_t\underline{\boldsymbol{\Pi}}_t\underline{\boldsymbol{Q}}_t'\underline{\boldsymbol{Q}}_t = \underline{\boldsymbol{\Pi}}_t, \tag{39}$$

have orthonormal columns

$$\widehat{\underline{\boldsymbol{Z}}}_t'\widehat{\underline{\boldsymbol{Z}}}_t = (\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t)'(\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t) = \underline{\boldsymbol{Q}}_t'(\widetilde{\underline{\boldsymbol{Z}}}_t'\widetilde{\underline{\boldsymbol{Z}}}_t)\underline{\boldsymbol{Q}}_t = \underline{\boldsymbol{Q}}_t'\boldsymbol{I}_k\underline{\boldsymbol{Q}}_t = \boldsymbol{I}_k, \tag{40}$$

and are orthogonal to the fundamental factor exposures (in whitened space)

$$\widehat{\underline{\boldsymbol{Z}}}_t'\underline{\boldsymbol{X}}_t = (\widetilde{\underline{\boldsymbol{Z}}}_t\underline{\boldsymbol{Q}}_t)'\underline{\boldsymbol{X}}_t = \underline{\boldsymbol{Q}}_t'(\widetilde{\underline{\boldsymbol{Z}}}_t'\underline{\boldsymbol{X}}_t) = \underline{\boldsymbol{Q}}_t'\mathbf{0} = \mathbf{0}. \tag{41}$$

If the final rotation $\underline{\boldsymbol{Q}}_t$ does not alter the statistical factors materially, then $\widehat{\underline{\boldsymbol{\Pi}}}_t \approx \underline{\boldsymbol{\Pi}}_t$. In practice, we use the time-series variances of the statistical factor returns $\widehat{\underline{f}}_{j,t+1} = \widehat{\underline{\boldsymbol{Z}}}_t'\underline{\boldsymbol{r}}_{t+1}$ to estimate the diagonal elements of $\widehat{\underline{\boldsymbol{\Pi}}}_t$, which may differ slightly from the eigenvalues in $\underline{\boldsymbol{\Pi}}_t$. This breaks the strict link between eigenvalues and factor variances but improves stability by anchoring variances directly to estimated statistical factor return series.

4.6 Hybrid Covariance in Whitened Space

If we assume that the factor covariance is block-diagonal, the hybrid covariance of whitened returns is

$$\underline{\boldsymbol{\Sigma}}_t = \begin{bmatrix} \underline{\boldsymbol{X}}_t & \underline{\widehat{\boldsymbol{Z}}}_t \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_t & \mathbf{0} \\ \mathbf{0} & \underline{\widehat{\boldsymbol{\Pi}}}_t \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{X}}_t & \underline{\widehat{\boldsymbol{Z}}}_t \end{bmatrix}' + \underline{\widehat{\boldsymbol{D}}}_t, \tag{42}$$

where $\widehat{\boldsymbol{\Omega}}_t = \mathrm{Cov}_t(\boldsymbol{b}_{t+1})$ is the covariance of fundamental factor returns, $\underline{\widehat{\boldsymbol{\Pi}}}_t$ are the factor return variances of the orthogonalized statistical factors, and $\underline{\widehat{\boldsymbol{D}}}_t$ is diagonal, containing the idiosyncratic variances after extracting the fundamental and statistical blocks.

This covariance matrix retains the standard factor model structure of exposures multiplying a factor covariance plus a diagonal specific variance. This structure can be useful for fast matrix inversion and is required by some portfolio analysis and portfolio construction tools.

Appendix A shows that scalar payoffs like factor returns are invariant to the return rotation. In our case, both OLS and GLS factor returns are consistent estimates of the true factor returns, even though their finite-sample estimates may differ. As a result, the covariance of factor returns $\widehat{\boldsymbol{\Omega}}_t$ does not require further adjustments once we have an estimate. This is convenient here because we can use the factor covariance we estimated for the factor returns in the fundamental model without modification. Consistent with our whitening assumptions, this is the sample covariance matrix of the GLS factor return estimates.

Although we relax the assumption that residual returns $\boldsymbol{e}_{t+1}$ are uncorrelated with each other, we continue to assume that they remain uncorrelated with the fundamental factor returns. This is a key justification for modeling additional structure in the fundamental residuals without disrupting the core interpretation of the fundamental factors.

In the hybrid risk model, we further orthogonalize the statistical factor exposures with respect to the fundamental exposures. This ensures that the statistical factors capture variation not already explained by the fundamental model. However, this orthogonality holds only in the space of asset exposures. It does not guarantee zero correlation between realized statistical and fundamental factor returns in finite samples.

In principle, if the statistical factor exposures $\underline{\widehat{\boldsymbol{Z}}}_t$ are estimated using only data available at time t, one could compute sample correlations between the corresponding statistical factor returns $\underline{\widehat{\boldsymbol{f}}}_{t+1} = \underline{\widehat{\boldsymbol{Z}}}_t' \underline{\boldsymbol{r}}_{t+1}$ and the fundamental factor returns $\widehat{\boldsymbol{b}}_{t+1}$. However, as emphasized by Miller (2006), such an analysis is rarely meaningful in practice. Statistical factors (e.g., principal

components of residuals) are rotation-invariant and may flip signs arbitrarily across time. These sign instabilities undermine the interpretability of multi-period factor returns and render correlation estimates inherently unreliable.

While the hybrid model construction implies zero correlation across fundamental and statistical factor returns and zero correlation among statistical factor returns, this is best viewed as a convenient modeling simplification, not a testable empirical restriction.

4.7 Mapping Back to Raw Returns

While it is convenient to estimate the covariance matrix in equation (42) in whitened space, we want a covariance matrix we can use with raw, observed returns. We find this covariance by transforming the covariance for whitened returns back to the raw return space.

The whitening is $\underline{\boldsymbol{r}}_{t+1} = \boldsymbol{\Gamma}_t^{1/2}\boldsymbol{r}_{t+1}$, so the raw-return covariance is

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Gamma}_t^{-1/2}\underline{\boldsymbol{\Sigma}}_t\boldsymbol{\Gamma}_t^{-1/2}, \tag{43}$$

which applies to raw, observed asset returns $\boldsymbol{r}_{t+1}$.

4.8 Covariance Estimates

We estimate $\widehat{\boldsymbol{\Omega}}_t$, $\widehat{\underline{\boldsymbol{\Pi}}}_t$, and $\widehat{\underline{\boldsymbol{D}}}_t$ in whitened space from time series of returns before assembling them into the overall covariance and the mapping back to raw return space. As shown in appendix A, fundamental factor returns are numerically identical in raw and whitened space when exposures are transformed consistently, so either set of return series can be used to estimate $\boldsymbol{\Omega}_t$. As before, we can apply exponential weighting to all of these covariance estimates.

The corresponding whitened hybrid residuals used to estimate $\widehat{\underline{\boldsymbol{D}}}_t$ are

$$\widehat{\underline{\boldsymbol{e}}}_{t+1} = \left(\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t} - \widehat{\underline{\boldsymbol{Z}}}_t\widehat{\underline{\boldsymbol{Z}}}_t'\right)\underline{\boldsymbol{r}}_{t+1}. \tag{44}$$

Because $\underline{\boldsymbol{P}}_{X,t}\widehat{\underline{\boldsymbol{Z}}}_t = \boldsymbol{0}$, the two projectors are orthogonal and commute, so the projector onto the complement of the combined factor space can be written equivalently using any of

$$\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t} - \widehat{\underline{\boldsymbol{Z}}}_t\widehat{\underline{\boldsymbol{Z}}}_t' = (\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t})(\boldsymbol{I}_N - \widehat{\underline{\boldsymbol{Z}}}_t\widehat{\underline{\boldsymbol{Z}}}_t') = (\boldsymbol{I}_N - \widehat{\underline{\boldsymbol{Z}}}_t\widehat{\underline{\boldsymbol{Z}}}_t')(\boldsymbol{I}_N - \underline{\boldsymbol{P}}_{X,t}). \tag{45}$$

We estimate $\widehat{\underline{\boldsymbol{D}}}_t$ by placing the sample variances of $\widehat{\underline{\boldsymbol{e}}}_{t+1}$ on the diagonal. Let the rolling statistical factor returns be

$$\widehat{\underline{\boldsymbol{f}}}_{t+1-s} = \widehat{\underline{\boldsymbol{Z}}}_{t-s}'\underline{\boldsymbol{r}}_{t+1-s}, \tag{46}$$

and define the constant-loadings series, with today's factor loadings applied to past returns, as

$$\underline{\widehat{f}}^{\star}_{t+1-s} = \underline{\widehat{\boldsymbol{Z}}}_t' \underline{\boldsymbol{r}}_{t+1-s}. \tag{47}$$

The rolling series use each date's contemporaneous statistical factor exposures, while the constant-loading series freezes today's exposures and applies them backwards in time. Because statistical exposures drift over time, the rolling series can "mix" nearby components (e.g., factor 3 migrating toward factor 2). For near-term risk forecasts, it may be preferable to estimate the diagonal $\underline{\widehat{\boldsymbol{\Pi}}}_t$ from the constant-loading series $\underline{\widehat{f}}^{\star}_{t+1-s}$, which measures the variance of today's statistical factors under recent return realizations.[7] If statistical factor exposures change slowly and we estimate the variances over relatively short sample periods, the constant-loading estimates may yield more stable short-term variance forecasts than the rolling series because they avoid mixing nearby components and yield variance forecasts aligned with the current loading structure.

We can shrink the k variance estimates toward their mean

$$\underline{\widehat{\pi}}_{j,t} = (1-\theta)\,\mathrm{Var}_t\left(\underline{\widehat{f}}^{\star}_{j}\right) + \theta\,\overline{\underline{\pi}}_t \tag{48}$$

$$\overline{\underline{\pi}}_t = \frac{1}{k}\sum_{j=1}^{k} \mathrm{Var}_t\left(\underline{\widehat{f}}^{\star}_{j}\right), \tag{49}$$

where the sample variances $\mathrm{Var}_t\left(\underline{\widehat{f}}^{\star}_{j}\right)$ can be exponentially weighted. A reasonable shrinkage intensity is

$$\theta = \frac{\nu}{1+\nu}, \qquad \text{with } \nu = \frac{k}{\tau}, \tag{50}$$

where τ is the effective number of observations defined earlier. This is the same linear shrinkage rule we applied to eigenvalues earlier, but here it is applied to the time-series variance estimates. Shrinkage reduces dispersion in the variance estimates, improving stability without altering the block-diagonal structure. As always, we can apply exponential weighting when computing these variances. In practice, mild floors/ceilings on $\underline{\widehat{\pi}}_{j,t}$ can further help numerical stability.

[7] These constant-loading series are not tradable because we did not know the factor weights on previous dates. They are forecasting devices. Although the constant-loading series fix the sign of the exposures over time, they are still not suitable for estimating off-diagonal elements in $\underline{\widehat{\boldsymbol{\Pi}}}_t$. The statistical factor covariance is diagonal by construction and the true elements are zero. Nonzero sample covariance estimates are noise.

5 Conclusion

By carefully combining a fundamental factor model with a statistical factor model applied to the residuals, we can improve the modeling of asset risk without interfering with the economic interpretation of the fundamental factors. The hybrid model retains the intuitive structure of the fundamental framework while allowing additional residual risk structure to be estimated.

The statistical components may significantly improve short-term risk estimation, especially in high-dimensional portfolios where the residual covariance from the fundamental model can be noisy and incomplete. However, since statistical exposures are unstable over time and arbitrarily signed, their projected returns should not be used for return attribution.

In practice, the hybrid model often serves as a stepping stone: residual structure captured statistically today may suggest directions for defining new fundamental factors tomorrow.

6 References

Bai, Jushan, and Serena Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191–221.

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 31, 307–327.

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

Connor, Gregory, and Robert A. Korajczyk, 1986, Performance measurement with the arbitrage pricing theory: A new framework for analysis, *Journal of Financial Economics* 15, 373–394.

Connor, Gregory, and Robert A. Korajczyk, 1993, A test for the number of factors in an approximate factor model, *Journal of Finance* 48, 1263–1291.

Engle, Robert F., 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* 50, 987–1008.

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

Hentschel, Ludger, 2024, Exponential weighting: Effects on precision, Working paper, Versor Investments, New York, NY.

Johnstone, Iain M., 2001, On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics* 29, 295–327.

Jones, Christopher S., 2001, Extracting factors from heteroskedastic asset returns, *Journal of Financial Economics* 62, 293–325.

J.P. Morgan/Reuters, 1996, *RiskMetrics – Technical Document*, New York, fourth edition.

Ledoit, Olivier, and Michael Wolf, 2015, Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions, *Journal of Multivariate Analysis* 139, 360–384.

Ledoit, Olivier, and Michael Wolf, 2020, Analytic nonlinear shrinkage of large-dimensional covariance matrices, *The Annals of Statistics* 48, 3043–3365.

Marchenko, Volodymyr A., and Leonid A. Pastur, 1967, Distribution of eigenvalues for some sets of random matrices, *Mathematics of the USSR-Sbornik* 1, 457–483.

Menchero, Jose, and Indrajit Mitra, 2008, The structure of hybrid factor models, *Journal of Investment Management* 6, 35–47.

Miller, Guy, 2006, Needles, haystacks, and hidden factors, *Journal of Portfolio Management* 32, 25–32.

A Raw and Whitened Factor Returns

We whiten raw returns $\boldsymbol{r}_{t+1}$ with a weighting matrix to find

$$\underline{\boldsymbol{r}}_{t+1} = \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{r}_{t+1}, \tag{51}$$

where the weighting matrix $\boldsymbol{\Gamma}_t$ is symmetric, positive definite, and often diagonal. This is a change in coordinates for the returns. Such a change may be advantageous for statistical precision or analytical convenience.

If raw returns have a factor covariance matrix

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

then whitened returns have a covariance matrix

$$\begin{aligned} \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{\Sigma}_t \boldsymbol{\Gamma}_t^{1/2} &= \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{X}_t \boldsymbol{\Omega}_t \boldsymbol{X}_t' \boldsymbol{\Gamma}_t^{1/2} + \boldsymbol{\Gamma}_t^{1/2} \boldsymbol{D}_t \boldsymbol{\Gamma}_t^{1/2} && (53) \\ &= \underline{\boldsymbol{X}}_t \boldsymbol{\Omega}_t \underline{\boldsymbol{X}}_t' + \underline{\boldsymbol{D}}_t. && (54) \end{aligned}$$

Here, the underlined variables are the whitened versions of the same variables in raw return space.

The same factor covariance $\boldsymbol{\Omega}_t$ that applies to raw returns also appears in whitened space, but now sandwiched by the whitened exposures $\underline{\boldsymbol{X}}_t$. This is because a one-period factor return is a scalar payoff $\boldsymbol{x}_t' \boldsymbol{r}_{t+1}$, and scalar payoffs are invariant to a change of coordinates in the return vector. Hence, factor returns – and their covariance matrix $\boldsymbol{\Omega}_t$ – are unaffected by whitening.

Both OLS and GLS provide consistent estimates of the true factor returns. In particular, there is no need to adjust the GLS estimate although we can think of it as the result of a regression in whitened space.

Of course, the same logic applies to the statistical factor returns and their covariance matrix. In practice, however, statistical factors are defined and estimated directly in whitened space. Outside of this space, the statistical exposures are not observed, so it is natural to treat their covariance as a whitened-space object.

By contrast, the whitened residual returns $\widehat{\underline{\boldsymbol{\varepsilon}}}_{t+1}$ and $\widehat{\underline{\boldsymbol{e}}}_{t+1}$ are different from their raw counterparts. They are vectors that live in the space of returns, not scalar payoffs that are linear functions of the returns. If we estimate the variances in whitened space, we must transform the estimates back to raw space.