A Appendix: Additional Results and Proofs

Additional Results on Alpha Hacking

We consider the situation where the researcher has in-sample data from time 1 to time T and an out-of-sample (oos) period from time $T+1$ to $T+T^{oos}$. The researcher may have used alpha-hacking during the in-sample period, but this does not affect the out-of-sample period. The researcher is interested in the posterior alpha based on the total evidence, in-sample and out-of-sample, which is useful for predicting factor performance in a future time period (that is, a time period that is out-of-sample relative to the existing out-of-sample period).

Proposition 6 (Out-of-sample alpha) *The posterior alpha based on an in-sample data from time 1 to T with alpha-hacking, and an out-of-sample period from $T+1$ to $T+T^{oos}$ is given by*

$$E(\alpha|\hat{\alpha},\hat{\alpha}^{oos}) = \kappa^{oos}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right) \tag{A.1}$$

where $w=\frac{\sigma^2/T^{oos}}{\bar{\sigma}^2/T+\sigma^2/T^{oos}}\in(0,1)$ is the relative weight on the in-sample period relative to the out-of-sample period, and $\kappa^{oos}=\frac{1}{1+1/(\tau^2([\bar{\sigma}^2/T]^{-1}+[\sigma^2/T^{oos}]^{-1}))}$ is a shrinkage parameter.

We see that, the more alpha hacking the researcher has done (higher $\bar{\sigma}$), the less weight we put on the in-sample period relative to the out-of-sample period. Further, the in-sample period has the non-proportional discounting due to alpha hacking ($\bar{\varepsilon}$), which we don't have for out-of-sample evidence.

So this result formalizes the idea that an in-sample backtest plus live performance is *not* the same as a longer backtest. For example, 10 years of backtest plus 10 years of live performance is more meaningful that 20 years of backtest with no live performance. The difference is that the oos-performance is free from alpha-hacking.

Proofs and Lemmas

The proofs make repeated use of the following well-known property of multivariate Normally distributed random variable. If x and y are multivariate Normal:

$$\begin{bmatrix}x\\y\end{bmatrix}\sim N\left(\begin{bmatrix}\mu_x\\\mu_y\end{bmatrix},\begin{bmatrix}\Sigma_{xx}&\Sigma_{yx}\\\Sigma_{xy}&\Sigma_{yy}\end{bmatrix}\right) \tag{A.2}$$

then the conditional distribution of x given y has the following Normal distribution:

$$x|y\sim N\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y),\,\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right) \tag{A.3}$$

The proofs also make use of the following two lemmas.

Lemma 1 *For random variables x,y,z, it holds that $E(\mathrm{Var}(x|y,z))\leq E(\mathrm{Var}(x|y))$ and, if the random variables are jointly normal, then $\mathrm{Var}(x|y,z)\leq\mathrm{Var}(x|y)$.*

Lemma 2 *Let A be an $N \times N$ matrix for which all diagonal elements equal a and all off-diagonal elements equal b, where $a \neq b$ and $a + b(N-1) \neq 0$. Then the inverse A^{-1} exists and is of the same form:*

$$A = \begin{bmatrix} a & & b \\ & \ddots & \\ b & & a \end{bmatrix} \quad A^{-1} = \begin{bmatrix} c & & d \\ & \ddots & \\ d & & c \end{bmatrix} \tag{A.4}$$

where $c = \frac{a+b(N-2)}{(a-b)(a+b(N-1))}$ *and* $d = \frac{-b}{(a-b)(a+b(N-1))}$.

Proof of Lemma 1. Using the definition of conditional variance, we have

$$E(\text{Var}(x|y,z)) = E(E(x^2|y,z)) - E([E(x|y,z)]^2) = E(x^2) - E([E(x|y,z)]^2)$$

Hence, using Jensen's inequality, we have

$$\begin{aligned} E(\text{Var}(x|y)) - E(\text{Var}(x|y,z)) =& E([E(x|y,z)]^2) - E([E(x|y)]^2) \\ =& E([E(x|y,z)]^2) - E([E(E(x|y,z)|y)]^2) \\ \geq& E([E(x|y,z)]^2) - E(E([E(x|y,z)]^2|y)) = 0 \end{aligned}$$

The result for normal distributions follows from the fact that normal conditional variances are non-stochastic, i.e., $\text{Var}(x|y) = E(\text{Var}(x|y))$. In this case, we can also characterize the extra drop in variance due to conditioning on z using its orthogonal component ε from the regression $z = a + by + \varepsilon$, using similar notation as (A.2):

$$\begin{aligned} \text{Var}(x|y,z) = \text{Var}(x|y,\varepsilon) =& \Sigma_{x,x} - \Sigma_{x,(y,\varepsilon)}\Sigma^{-1}_{(y,\varepsilon),(y,\varepsilon)}\Sigma_{(y,\varepsilon),x} \\ =& \Sigma_{x,x} - \Sigma_{x,y}\Sigma^{-1}_{y,y}\Sigma_{y,x} - \Sigma_{x,\varepsilon}\Sigma^{-1}_{\varepsilon,\varepsilon}\Sigma_{\varepsilon,x} = \text{Var}(x|y) - \Sigma_{x,\varepsilon}\Sigma^{-1}_{\varepsilon,\varepsilon}\Sigma_{\varepsilon,x} \end{aligned}$$

■

Proof of Lemma 2. The proof follows from inspection: The product of A and its proposed inverse clearly has the same form as A with diagonal elements

$$ac + bd(I-1) = \frac{a(a+b(N-2)) - b^2(N-1)}{(a-b)(a+b(N-1))} = \frac{a^2 + ab(N-1) - ab - b^2(N-1)}{(a-b)(a+b(N-1))} = 1$$

and off-diagonal elements

$$ad + bc + bd(N-2) = \frac{-ab + b(a+b(N-2)) - b^2(N-2)}{(a-b)^2(a+b(N-1))^2} = 0$$

In other words, AA^{-1} equals the identity, proving the result. ■

Proof of Equations (4)–(6). The posterior distribution of the true alpha given the observed

factor return is computed using (A.3). The conditional mean is

$$E(\alpha|\hat{\alpha}) = 0 + \frac{\text{Cov}(\alpha,\hat{\alpha})}{\text{Var}(\hat{\alpha})}(\hat{\alpha} - 0) = \frac{\tau^2}{\tau^2 + \sigma^2/T}\hat{\alpha} = \ \kappa\hat{\alpha}$$

where κ is given by (5) and the posterior variance is

$$\text{Var}(\alpha|\hat{\alpha}) = \text{Var}(\alpha) - \frac{(\text{Cov}(\alpha,\hat{\alpha}))^2}{\text{Var}(\hat{\alpha})} = \tau^2 - \tau^2\frac{\tau^2}{\tau^2 + \sigma^2/T} = \frac{\tau^2\sigma^2/T}{\tau^2 + \sigma^2/T} = \kappa\frac{\sigma^2}{T}$$

■

Proof of Proposition 1. The posterior alpha with alpha-hacking is given via (A.3) as

$$E(\alpha|\hat{\alpha}) = 0 + \frac{\text{Cov}(\alpha,\hat{\alpha})}{\text{Var}(\hat{\alpha})}(\hat{\alpha} - E(\hat{\alpha})) = \frac{\tau^2}{\tau^2 + \bar{\sigma}^2/T}(\hat{\alpha} - \bar{\varepsilon}) = -\kappa_0 + \kappa^{\text{hacking}}\hat{\alpha}$$

where $\kappa^{\text{hacking}} = \frac{1}{1+\frac{\bar{\sigma}^2}{\tau^2 T}}$, $\kappa_0 = \kappa^{\text{hacking}}\bar{\varepsilon} \geq 0$, and $\kappa^{\text{hacking}} \leq \kappa$ because $\bar{\sigma} \geq \sigma$. ■

Proof of Proposition 2. The posterior mean given $\hat{\alpha}$ and $\hat{\alpha}^g$ is computed via (A.3) as

$$\begin{aligned}
E(\alpha|\hat{\alpha},\hat{\alpha}^g) &= \begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\sigma_T^2 & \tau^2+\rho\sigma_T^2\\ \tau^2+\rho\sigma_T^2 & \tau^2+\sigma_T^2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\alpha}\\ \hat{\alpha}^g\end{bmatrix}\\
&= \frac{1}{\det}\begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\sigma_T^2 & -(\tau^2+\rho\sigma_T^2)\\ -(\tau^2+\rho\sigma_T^2) & \tau^2+\sigma_T^2\end{bmatrix}\begin{bmatrix}\hat{\alpha}\\ \hat{\alpha}^g\end{bmatrix}\\
&= \frac{\tau^2(1-\rho)\sigma_T^2}{\det}(\hat{\alpha}+\hat{\alpha}^g)\\
&= \frac{\tau^2(1-\rho)}{\sigma_T^2(1-\rho)(1+\rho)+2\tau^2(1-\rho)}(\hat{\alpha}+\hat{\alpha}^g)\\
&= \ \kappa^g\left(\frac{1}{2}\hat{\alpha}+\frac{1}{2}\hat{\alpha}^g\right)
\end{aligned}$$

using the notation $\sigma_T^2 = \sigma^2/T$ and

$$\det = (\tau^2+\sigma_T^2)^2 - (\tau^2+\rho\sigma_T^2)^2 = \sigma_T^2[\sigma_T^2(1-\rho^2) + 2\tau^2(1-\rho)].$$

The global shrinkage parameter κ^g is in $[\kappa, 1]$ and decreases with the correlation ρ, attaining the minimum value, $\kappa^g = \kappa$, when $\rho = 1$ as is clearly seen from (12).

The result about the posterior variance follows from Lemma 1. ■

Proof of Proposition 3. The prior joint distribution of the true and estimated alphas is

given by the following expression, where we focus on factor 1 without loss of generality:

$$\begin{bmatrix} \alpha^1 \\ \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_c^2+\tau_w^2 & \tau_c^2+\tau_w^2 & \tau_c^2 & \cdots & \tau_c^2 \\ \tau_c^2+\tau_w^2 & \tau_c^2+\tau_w^2+\sigma^2/T & & & \tau_c^2+\rho\sigma^2/T \\ \tau_c^2 & & & & \\ \vdots & & & \ddots & \\ \tau_c^2 & \tau_c^2+\rho\sigma^2/T & & & \tau_c^2+\tau_w^2+\sigma^2/T \end{bmatrix} \right) \tag{A.5}$$

The posterior alpha of factor 1 is therefore normally distributed with a mean derived using the standard formula for conditional normal distributions (A.3):

$$E(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) = \begin{bmatrix} \tau_c^2+\tau_w^2 \\ \tau_c^2 \\ \vdots \\ \tau_c^2 \end{bmatrix}^\top \begin{bmatrix} \tau_c^2+\tau_w^2+\sigma^2/T & & \tau_c^2+\rho\sigma^2/T \\ & \ddots & \\ \tau_c^2+\rho\sigma^2/T & & \tau_c^2+\tau_w^2+\sigma^2/T \end{bmatrix}^{-1} \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix}$$

We next use Lemma 2 and its notation, i.e., $a = \tau_c^2+\tau_w^2+\sigma^2/T$, $b = \tau_c^2+\rho\sigma^2/T$, and c', d are defined accordingly, where we use the notation c' to avoid confusion with the c in equation (14). This application of Lemma 2 yields

$$\begin{aligned}
E(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) &= \begin{bmatrix} \tau_c^2+\tau_w^2 \\ \tau_c^2 \\ \vdots \\ \tau_c^2 \end{bmatrix}^\top \begin{bmatrix} c' & & d \\ & \ddots & \\ d & & c' \end{bmatrix} \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \\
&= \begin{bmatrix} \tau_c^2(c'+d(N-1))+\tau_w^2 c' \\ \tau_c^2(c'+d(N-1))+\tau_w^2 d \\ \vdots \\ \tau_c^2(c'+d(N-1))+\tau_w^2 d \end{bmatrix}^\top \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \\
&= (\tau_c^2(c'+d(N-1))+\tau_w^2 d)N\hat{\alpha}^\cdot + \tau_w^2(c'-d)\hat{\alpha}^1 \\
&= (\tau_c^2\frac{N}{a+b(N-1)} - \tau_w^2\frac{bN}{(a-b)(a+b(N-1))})\hat{\alpha}^\cdot + \tau_w^2\frac{1}{a-b}\hat{\alpha}^1 \\
&= \frac{\tau_c^2}{b+\frac{a-b}{N}}\,\hat{\alpha}^\cdot + \frac{\tau_w^2}{a-b}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{a-b}{bN}}\,\hat{\alpha}^\cdot\right) \\
&= \frac{\tau_c^2}{\tau_c^2+\rho\sigma^2/T+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{N}}\,\hat{\alpha}^\cdot + \frac{\tau_w^2}{\tau_w^2+(1-\rho)\sigma^2/T}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\,\hat{\alpha}^\cdot\right) \\
&= \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{\tau_c^2 N}}\,\hat{\alpha}^\cdot + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\,\hat{\alpha}^\cdot\right)
\end{aligned}$$

The posterior has conditional variance

$$\begin{aligned}
\mathrm{Var}(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) =& \tau_c^2+\tau_w^2-\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}^\top\begin{bmatrix}c' & & d\\ & \ddots & \\ d & & c'\end{bmatrix}\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}\\
=& \tau_c^2+\tau_w^2-\begin{bmatrix}\tau_c^2(c'+d(N-1))+\tau_w^2c'\\ \tau_c^2(c'+d(N-1))+\tau_w^2d\\ \vdots\\ \tau_c^2(c'+d(N-1))+\tau_w^2d\end{bmatrix}^\top\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}\\
=& \tau_c^2+\tau_w^2-(\tau_c^2(c'+d(N-1))+\tau_w^2c')(\tau_c^2+\tau_w^2)\\
& -(\tau_c^2(c'+d(N-1))+\tau_w^2d)\tau_c^2(N-1)\\
\to& \tau_c^2+\tau_w^2-(\tau_c^2(\frac{1}{a-b}-\frac{1}{a-b})+\tau_w^2\frac{1}{a-b})(\tau_c^2+\tau_w^2)\\
& -(\tau_c^2\frac{1}{b}-\tau_w^2\frac{1}{a-b})\tau_c^2\\
=& \tau_c^2+\tau_w^2-\left(\tau_w^4\frac{1}{a-b}+\tau_c^4\frac{1}{b}\right)\\
=& \tau_c^2+\tau_w^2-\left(\frac{\tau_w^4}{\tau_w^2+(1-\rho)\sigma^2/T}+\frac{\tau_c^4}{\tau_c^2+\rho\sigma^2/T}\right)
\end{aligned}$$

The last results follow from Lemma 1. ■

Proof of Proposition 4. We write the joint prior distribution of true and observed alphas in the multi-level hierarchical model as

$$\begin{pmatrix}\alpha\\ \hat{\alpha}\end{pmatrix}\sim N\left(\alpha^0\,1_{2NK},\begin{pmatrix}\Omega & \Omega\\ \Omega & \Omega+\Sigma/T\end{pmatrix}\right) \tag{A.6}$$

The posterior mean vector of true alphas is computed via (A.3):

$$\begin{aligned}
E(\alpha|\hat{\alpha}) &= 1_{NK}\alpha_0+\Omega\left(\Omega+\Sigma/T\right)^{-1}\left(\hat{\alpha}-1_{NK}\alpha_0\right)\\
&= \left(\Omega^{-1}+T\Sigma^{-1}\right)^{-1}\left(\Omega^{-1}1_{NK}\alpha_0+T\Sigma^{-1}\hat{\alpha}\right)
\end{aligned}$$

using that $(\Omega+\Sigma/T)^{-1}=\Omega^{-1}-\Omega^{-1}\left(\Omega^{-1}+T\Sigma^{-1}\right)^{-1}\Omega^{-1}$ by the Woodbury matrix identity. The posterior variance is computed similarly via (A.3) and the same application of the Woodbury matrix identity as

$$\mathrm{Var}(\alpha|\hat{\alpha})=\Omega-\Omega\left(\Omega+\Sigma/T\right)^{-1}\Omega=\left(\Omega^{-1}+T\Sigma^{-1}\right)^{-1}.$$

■

Proof of Proposition 5. Based on the definition of the Bayesian FDR, we have:

$$\begin{aligned}
\text{FDR}^{\text{Bayes}} &= E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}}\middle|\hat{\alpha}^1,\ldots,\hat{\alpha}^N\right) \\
&= \frac{1}{\sum_i 1_{\{i \text{ discovery}\}}} E\left(\sum_i 1_{\{i \text{ false discovery}\}}\middle|\hat{\alpha}^1,\ldots,\hat{\alpha}^N\right) \\
&= \frac{1}{\sum_i 1_{\{i \text{ discovery}\}}} \sum_i Pr(i \text{ false discovery}|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) \\
&= \frac{1}{\#\text{discoveries}} \sum_{i \text{ discovery}} p\text{-}val_i^{Bayes} \\
&\leq 2.5\%
\end{aligned} \tag{A.7}$$

■

Proof of Proposition 6. The posterior mean alpha is

$$\begin{aligned}
E(\alpha|\hat{\alpha},\hat{\alpha}^{oos}) &= \begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\bar{\sigma}_T^2 & \tau^2 \\ \tau^2 & \tau^2+\sigma_{oos}^2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\alpha}-\bar{\varepsilon} \\ \hat{\alpha}^{oos}\end{bmatrix} \\
&= \frac{1}{\det}\begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\sigma_{oos}^2 & -\tau^2 \\ -\tau^2 & \tau^2+\bar{\sigma}_T^2\end{bmatrix}\begin{bmatrix}\hat{\alpha}-\bar{\varepsilon} \\ \hat{\alpha}^{oos}\end{bmatrix} \\
&= \frac{\tau^2}{\det}\left(\sigma_{oos}^2(\hat{\alpha}-\bar{\varepsilon})+\bar{\sigma}_T^2\hat{\alpha}^g\right) \\
&= \frac{\tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)}{\tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)+\bar{\sigma}_T^2\sigma_{oos}^2}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right) \\
&= \frac{\tau^2}{\tau^2+\bar{\sigma}_T^2\sigma_{oos}^2/(\bar{\sigma}_T^2+\sigma_{oos}^2)}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right) \\
&= \frac{1}{1+\frac{1}{\tau^2(\bar{\sigma}_T^{-2}+\sigma_{oos}^{-2})}}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right)
\end{aligned}$$

using the notation $\bar{\sigma}_T^2 = \bar{\sigma}^2/T$, $\sigma_{oos}^2 = \sigma^2/T^{oos}$, and

$$\det = (\tau^2+\bar{\sigma}_T^2)(\tau^2+\sigma_{oos}^2)-\tau^4 = \tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)+\bar{\sigma}_T^2\sigma_{oos}^2.$$

■

Accounting for Unobserved Factors

Harvey et al. (2016) provides a framework to estimate the total number of factors researchers have tried. The framework is based on t-statistics of published factors and estimation framework to determine the number of unobserved factors.

One set of simulations is constructed to match the baseline scenario of (Harvey et al., 2016, Table 5.A, row 1), which estimates that researchers have tried $M = 1,297$ factors, of which 39.6% of have zero alpha and another is based on the more conservative scenario

of (Harvey et al., 2016, Table 5.B, row 1), which implies that researchers have tried 2458 factors, of which 68.3% have zero alpha. Harvey et al. (2016) states that "the average annual Sharpe ratio for these [true] factors is 0.44."

To incorporate these unobserved factors into our framework, we proceed as follows for the baseline scenario. We simulate a total of 1,300 factors in 26 clusters of 50 factors per cluster. We let all factors in 10 clusters have true alphas equal to zero while the remaining clusters have non-zero true alphas. For each of the clusters with non-zero alphas, we set the cluster alpha to $c^j = 0.44 \times 10\%/12$ so that the monthly abnormal return corresponds to an annual Sharpe ratio of 0.44 given the annual volatility of 10%. Finally, we draw each factor's true alpha from $\alpha^i \sim N(c^j, \tau_w^2)$, and then simulate 68 years of monthly returns with within-cluster correlation of 0.5 and 0 otherwise. Finally, we estimate prior parameters τ using this data with the same method that we used on the observed data. We repeat this simulation process and compute the average τ_c, which is interpreted as a value that accounts for unobserved factors of the form implied by Harvey et al. (2016). We note that we are implicitly assuming that the unobserved factors belong to different clusters, such that observing new poor performing factors would lead to more shrinkage toward zero via a lower τ_c, but not via different cluster mean returns.

Similarly for the conservative scenario, we simulate a total of 2500 factors in 50 clusters of 50 factors per cluster. We let all factors in 16 clusters have true alphas equal to zero while the remaining clusters have non-zero true alphas as described above.

B Empirical Bayes Estimation

For convenient reference, we restate the multi-level hierarchical model of Section 1. For a factor i in cluster j and corresponding to signal n, the factor is

$$f_t^i = \alpha^i + \beta^i r_t^m + \varepsilon_t^i$$

with

$$\alpha^i = \alpha^o + c^j + s^n + w^i$$

where the alpha components are $\alpha^o = 0$, $c^j \sim N(0, \tau_c^2)$, $s^n \sim N(0, \tau_s^2)$, and $w^i \sim N(0, \tau_w^2)$. We write alpha in vector form as

$$\alpha = \alpha^o \, 1_{NK} + Mc + Zs + w \tag{B.1}$$

where $\alpha = (\alpha^1, \ldots, \alpha^{NK})'$, $c = (c^1, \ldots, c^J)'$, $s = (s^1, \ldots, s^N)'$, $w = (w^1, \ldots, w^{NK})'$, M is the $NK \times J$ matrix of cluster memberships, and Z is the $NK \times N$ matrix indicating the characteristic that factor i is based on. Given the hyperparameters $(\alpha^0, \tau_c, \tau_s, \tau_w)$, the prior mean and covariance matrix of alphas are

$$E[\alpha] = 0, \quad \Omega \equiv \mathrm{Var}(\alpha) = MM'\tau_c^2 + ZZ'\tau_s^2 + I_{NK}\tau_w^2.$$

The vector of return shocks is $\varepsilon_t = (\varepsilon_t^1, \ldots, \varepsilon_t^{NK})'$ which is distributed $\varepsilon_t \sim N(0, \Sigma)$.

Given this structure, we estimate the model as follows. The vector of factor returns $f_t = (f_t^1, ..., f_t^{NK})'$ has marginal likelihood—that is, after integrating out the uncertain alpha components—that is distributed as

$$f_t \sim N(0, [\Omega + \Sigma])$$

or, equivalently (treating CAPM betas as known), the estimated alphas are distributed[43]

$$\hat{\alpha} \sim N(0, [\Omega + \Sigma/T]).$$

The matrices Z and M are given by the factor definition and cluster assignment (Table C.4), respectively. We use a plug-in estimate of the factor CAPM-residual return covariance matrix, denoted $\hat{\Sigma}$ (discussed below). Finally, given $\hat{\Sigma}$, Z, and M, we estimate the hyperparameters of the prior distribution, (τ_c, τ_s, τ_w) via MLE based on the marginal likelihood.

This estimation approach is an example of the empirical Bayes method. It approximates the fully Bayesian posterior calculation (which requires integrating over a hyperprior distribution of hyperparameters, usually an onerous calculation) by setting the hyperparameters to their most likely values based on the marginal likelihood. It is particularly well suited to hierarchical Bayesian models in which parameters for individual observations share some common structure, so that the realized heterogeneity across individual is informative about sensible values for the hyperparameters of the prior. Our model and estimation approach implementation is a minor variation on Bayesian hierarchical normal mean models that are common in Bayesian statistics (textbook treatments include Efron, 2012; Gelman et al., 2013; Maritz, 2018). We conduct sensitivity analysis to ensure that our results are robust to a wide range of hyperparameters (see Figure C.4). Also, we note that our EB methodology is more easily replicable than a full-Bayesian setting with additional hyperpriors as EB relies on a closed-form Bayesian updating rather than a numerical integration.

To ensure cross-sectional stationarity, we scale each factor such that their monthly idiosyncratic volatility is $10\%/\sqrt{12}$ (i.e., 10% annualized). To construct a plug-in estimate of the factor residual return covariance matrix, denoted $\hat{\Sigma}$, we face two main empirical challenges. First, the sample covariance is poorly behaved due the relatively large number of factors compared to the number of time series observations. Second, we have an unbalanced panel because different factors come online at different points in time. To address the first challenge, we impose a block equicorrelation structure on Σ based on factors' cluster membership.[44] The correlation between factors in clusters i and j is estimated as the average correlation among all pairs such that one factor is in cluster i and the other is in j. In our global analyses, blocks correspond to region-cluster pairs. To address unbalancedness, we use the bootstrap. In particular, we generate 10,000 bootstrap samples that resample rows of the unbalanced factor return dataset. Each bootstrap sample is, therefore, also unbalanced, and we use this to produce a distribution of alpha estimates. From this we calculate $\hat{\Sigma}/T$

[43] We abstract from uncertainty in CAPM betas to emphasize the Bayesian updating of alphas. Our conclusions are qualitatively insensitive to accounting for beta uncertainty.

[44] As advocated by Engle and Kelly (2012) and Elton et al. (2006), block equicorrelation constrains all pairs of factors in the same block to share a single correlation parameter, and likewise for cross-block correlations. This stabilizes covariance matrix estimates by dramatically reducing the parameterization of the correlation matrix, while leaving the individual variance estimates unchanged.

as the covariance of alphas across bootstrap samples (imposing the block equicorrelation structure).

Table B.1 shows the estimated hyperparameters across different samples. While most of our analysis of based on these full-sample estimates, we also consider rolling-estimates of when considering out-of-sample evidence as seen in Figure C.7.

Sample	τ_c	τ_w	τ_s
USA	0.374	0.209	
Developed	0.263	0.183	
Emerging	0.307	0.233	
USA, Developed & Emerging	0.301	0.189	0.091
USA - Mega	0.272	0.152	
USA - Large	0.332	0.179	
USA - Small	0.462	0.264	
USA - Micro	0.501	0.337	
USA - Nano	0.535	0.350	

Table B.1: Hyperparameters of the prior distribution estimated by maximum likelihood. Here, τ_c is the estimated dispersion in cluster alphas (e.g., the dispersion in the alpha of the value cluster alpha, momentum cluster, and so on). When we estimate a single region, τ_w is the idiosyncratic dispersion of alphas within each cluster. When we jointly estimate several regions, then τ_s is the estimated dispersion in alphas across signals within each cluster, and τ_w is the estimated idiosyncratic dispersion in alphas for factors identified by their signal and region.

Internet Appendix

C Additional Exhibits

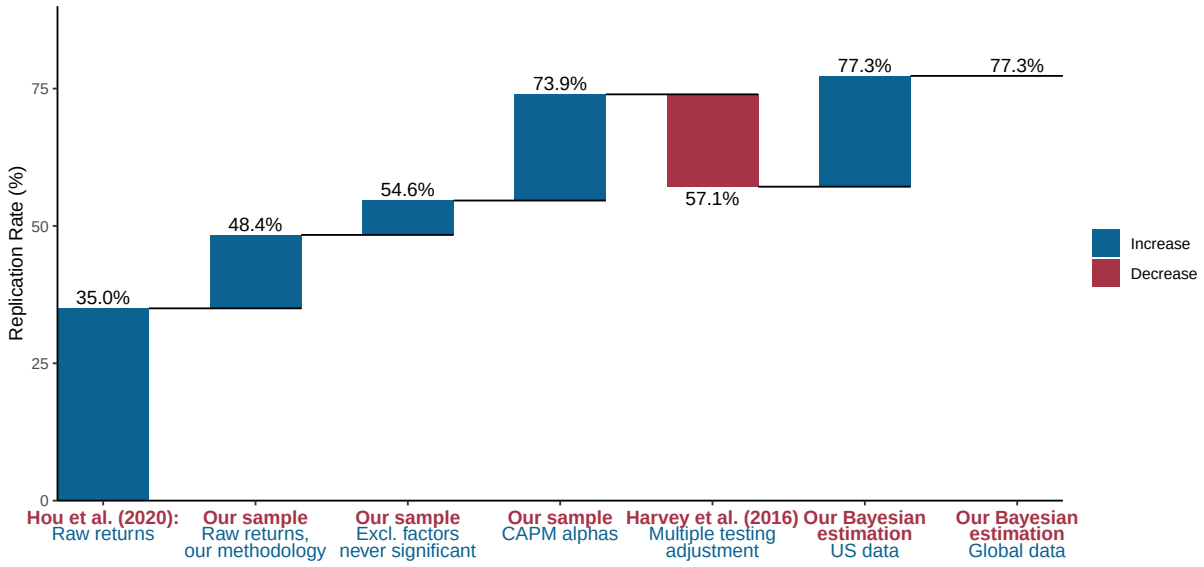

Figure C.1: Replication Rates Versus the Literature (Uncapped Value-weighting)

Note: This figure reproduces the analysis of Figure 1 using uncapped value weights to construct factors.

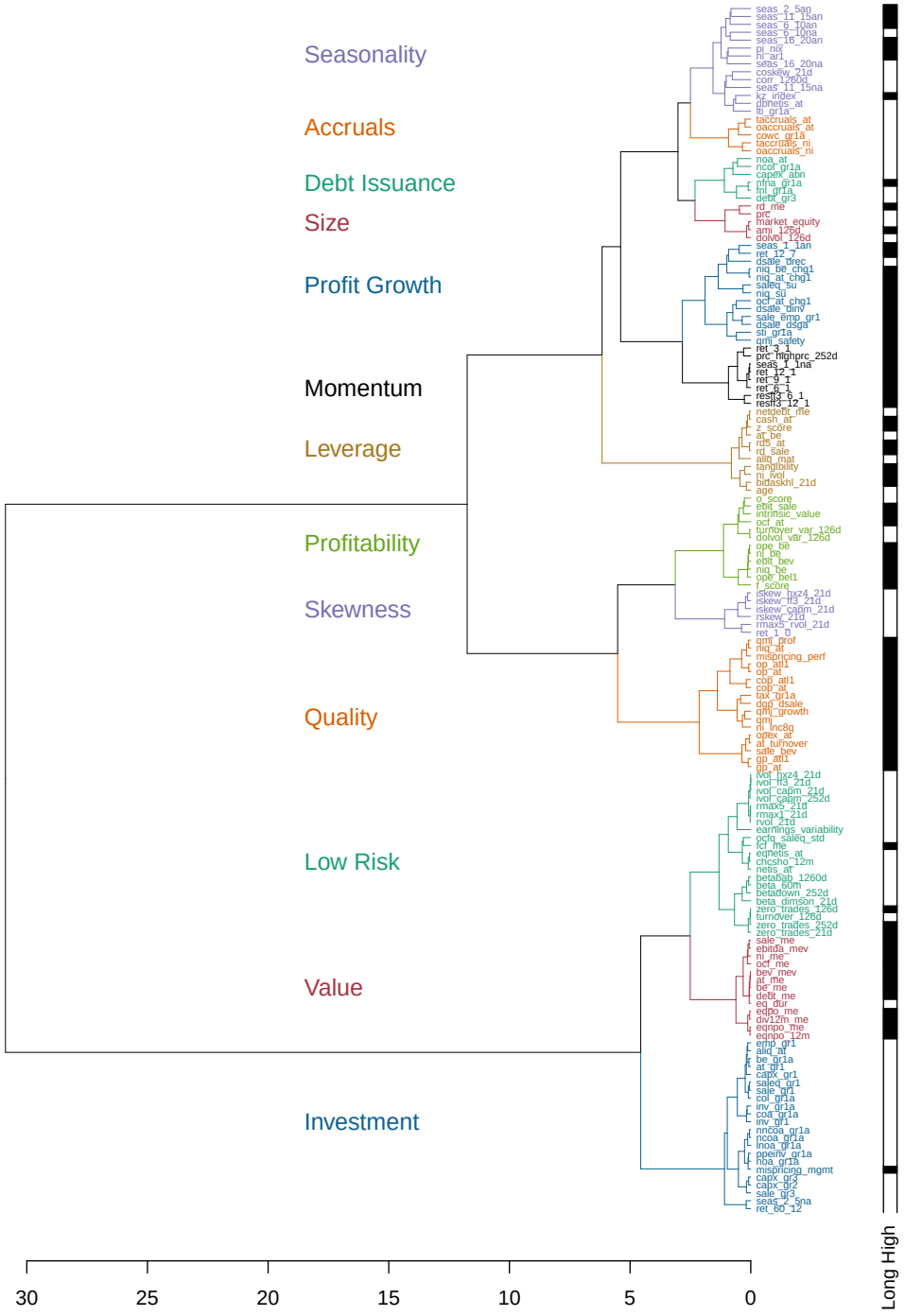

Figure C.2: Clustering Factors into Themes

Note: This figures shows a hierarchical clustering of all factors into 15 themes using the sample of US stocks from 1975-2019. Long high indicates whether the factor is long stocks with a high value of the underlying characteristic.

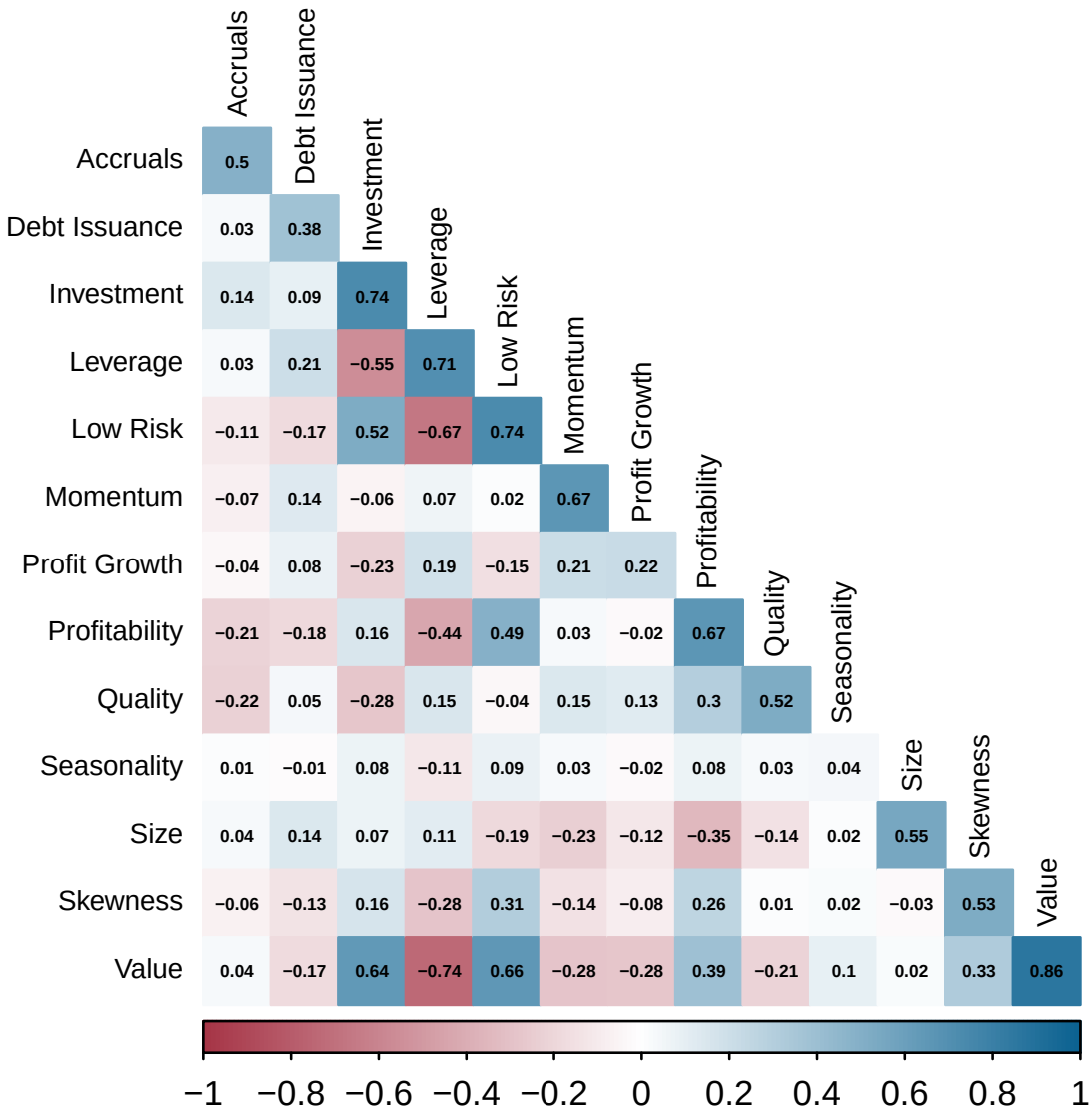

Figure C.3: Factor Theme Correlations

Note: This figure shows the average pairwise Pearson correlation between factors from different clusters (off diagonal elements) or between factors in the same cluster (diagonal elements), using data on US stocks during the 1975-2019 period.

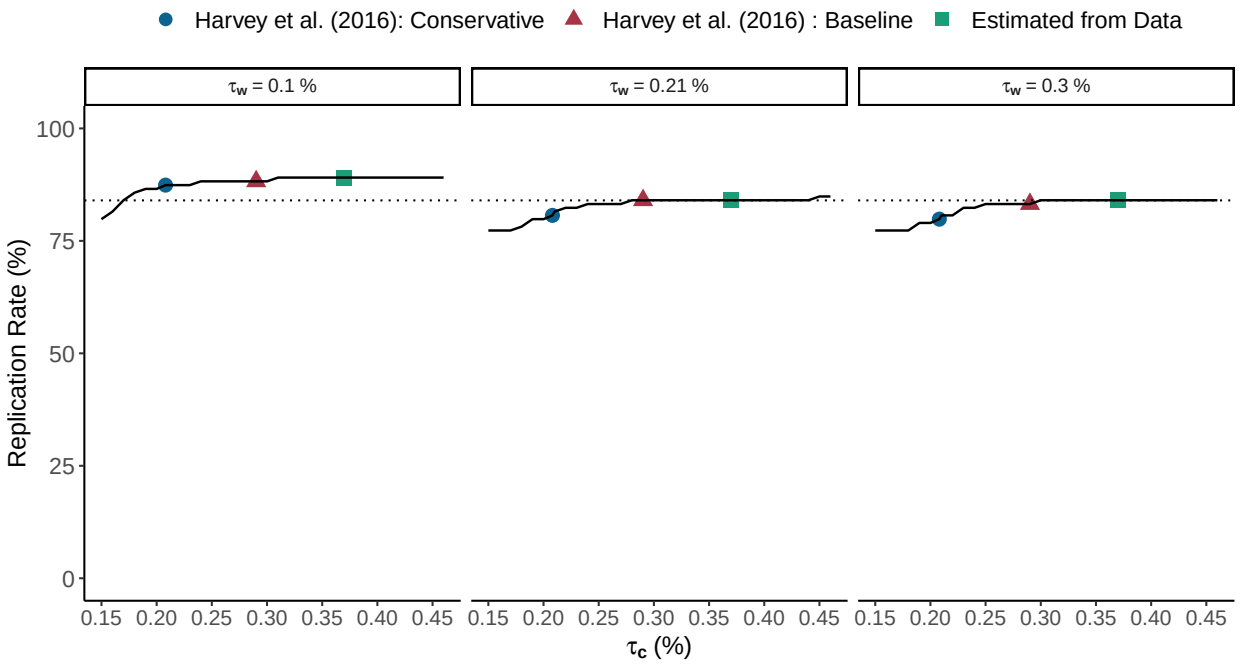

Figure C.4: Replication Rate with Prior Estimated in Light of Unobserved Factors

Note: The figure shows how the replication rate in the US varies when changing the τ_c and τ_w parameter. The dotted line shows our replication rate of 84%. The data estimate of τ_w is 0.21%. The green square, highlights the value estimated in the data $\tau_c = 0.37\%$. The red triangle and the blue circle highlights values that are found by estimating the empirical Bayes model according to assumptions about unobserved factors from Harvey et al. (2016). The values are $\tau_c = 0.29\%$ in the baseline scenario and $\tau_c = 0.21\%$ in the conservative scenario. A description of this approach can be found in the appendix, section A.

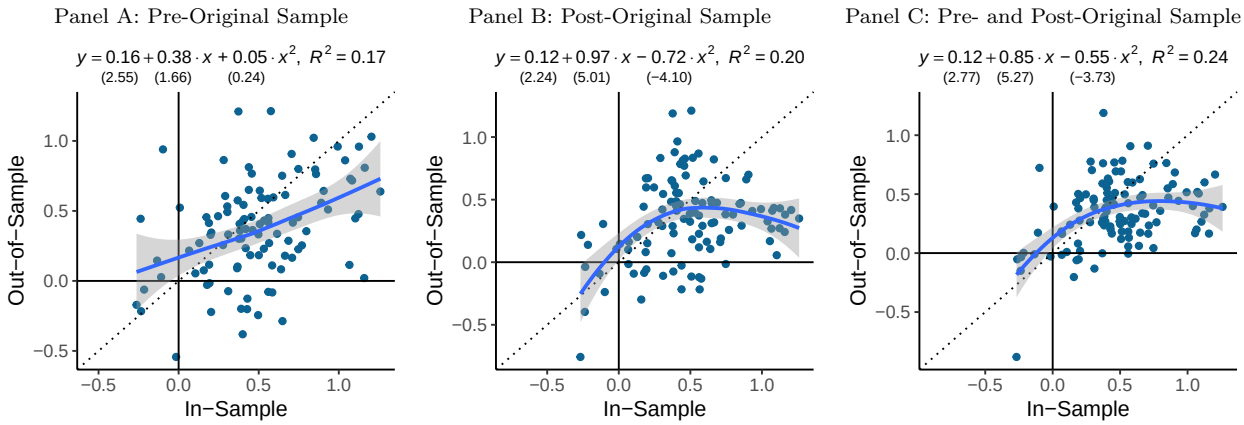

Figure C.5: In-Sample versus Out-of-Sample Alphas for US Factors

Note: The figure plots OLS alphas for US factors during the in-sample period (i.e., the period studied in the original publication) versus out-of-sample alphas. In Panel A, out-of-sample is the time period before the in-sample period. In Panel B, out-of-sample is the time period before the in-sample period. In Panel C, out-of-sample includes both the time period before and after the in-sample period. We require at least five years of out-of-sample data for a factor to be included, amounting to 103, 109 and 113 factors in panel A, B and C. The figure also reports an OLS regression of out-of-sample alphas on in-sample alphas and in-sample alphas squared. The blue line is a local polynomial regression fit where observations are weighted by their vicinity to the point on the x-axis. The shaded area is 95% confidence bands. The dotted line is the 45° line.

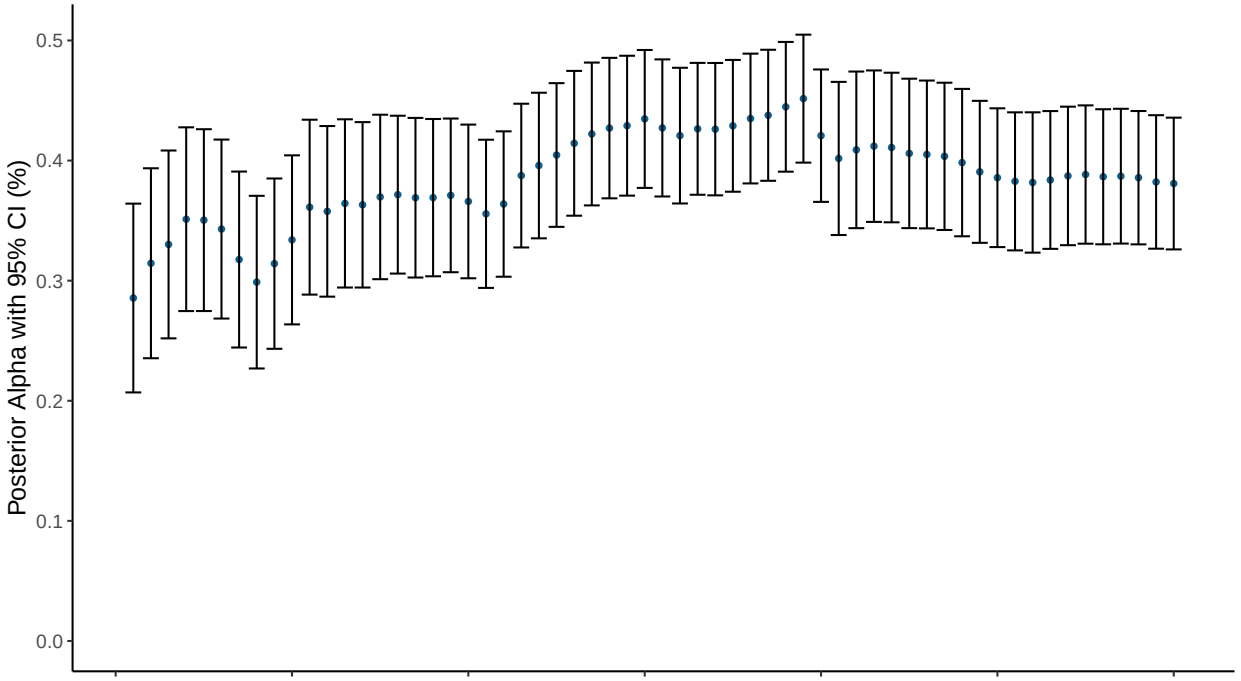

Figure C.6: US Factor Alpha Posterior Distribution Over Time

Note: The figure reports the average 95% posterior confidence interval for US factors based on EB posteriors re-estimated in December each year. In contrast to figure 8, we re-estimate τ_c and τ_w at each point in time. Figure C.7 shows how the estimated taus evolve over time.

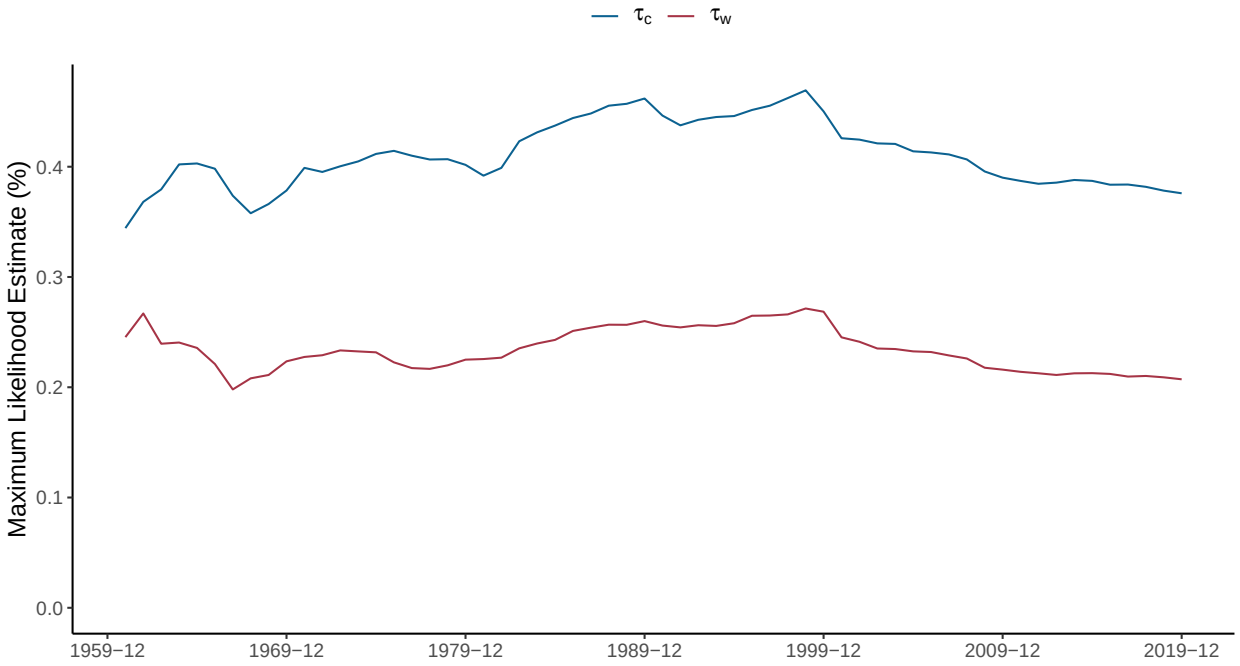

Figure C.7: US Factor Alpha Posterior Distribution Over Time

Note: The figure reports the τ_c and τ_w used in figure C.6.

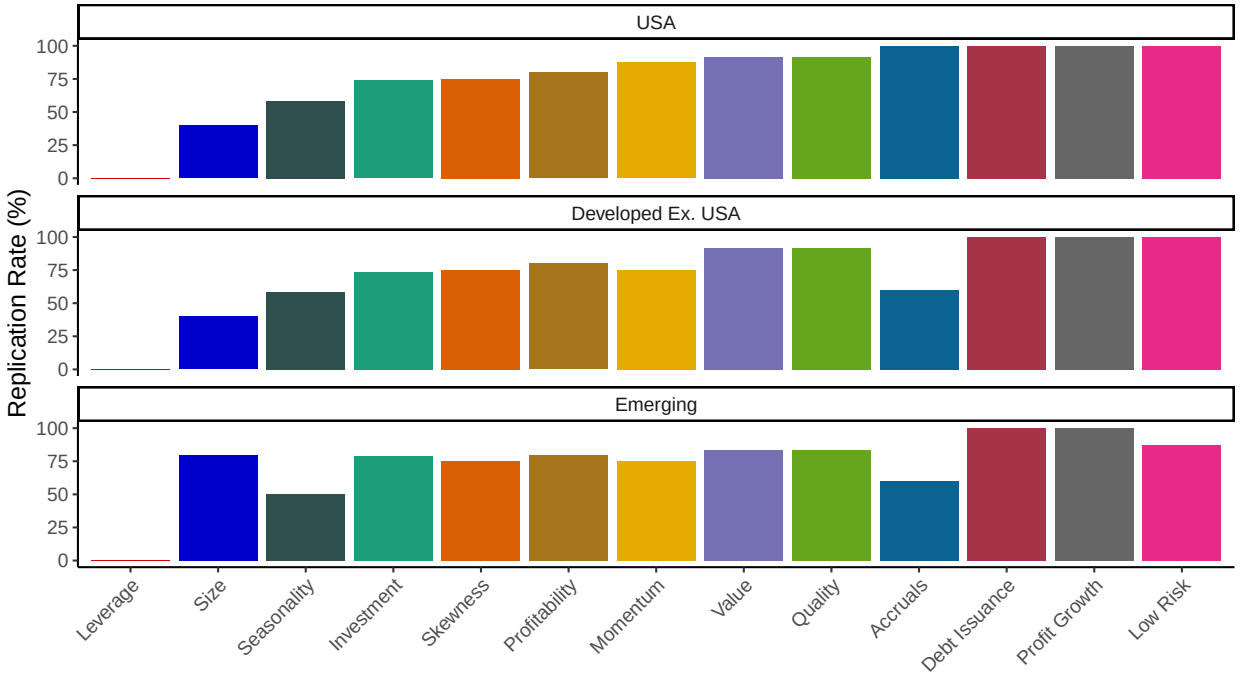

Figure C.8: Replication Rates across Regions by Cluster

Note: Share of factors within each cluster where the 95% posterior intervals does not include zero.

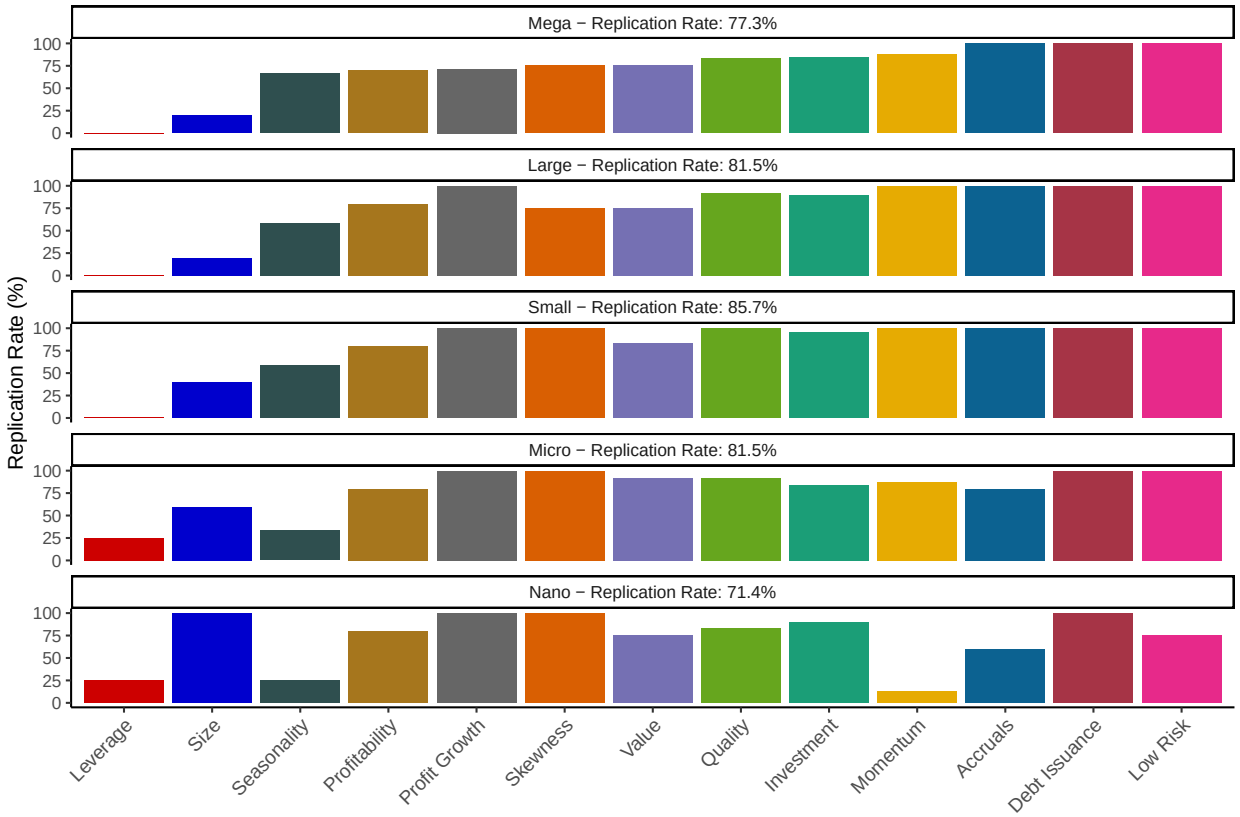

Figure C.9: Replication Rates across Size Groups by Cluster

Note: The figure shows replication rates for US factors created within a size group using rank weights. Mega stocks have a market cap higher than the 80th percentile of NYSE stocks, large stocks are between the 80th and 50th percentile, small stocks are between the 50th and 20th percentile, micro stocks are between the 20th and 1st percentile and nano stocks have a market cap below the 1st percentile of NYSE stocks.

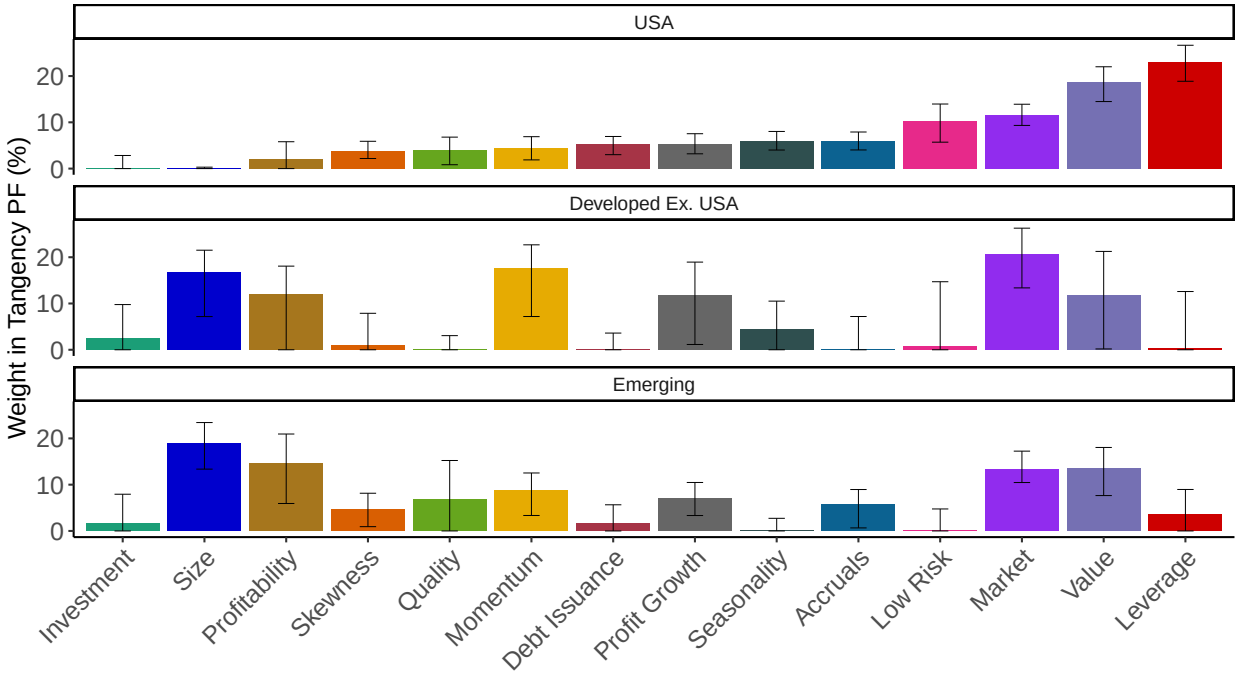

Figure C.10: Tangency Portfolio Weights across Regions

Note: Within each region, we compute the cluster return as the equal weighted return of all factors with data available at a given point in time. We further add the regional market return. We estimate the tangency weights following the method of Britten-Jones (1999) with a non-negativity constraint. The error bars are the 90% confidence intervals based on 10,000 bootstrap samples and the percentile method. The data starts in 1952 for the US, 1987 for Developed ex. US and 1994 for Emerging.

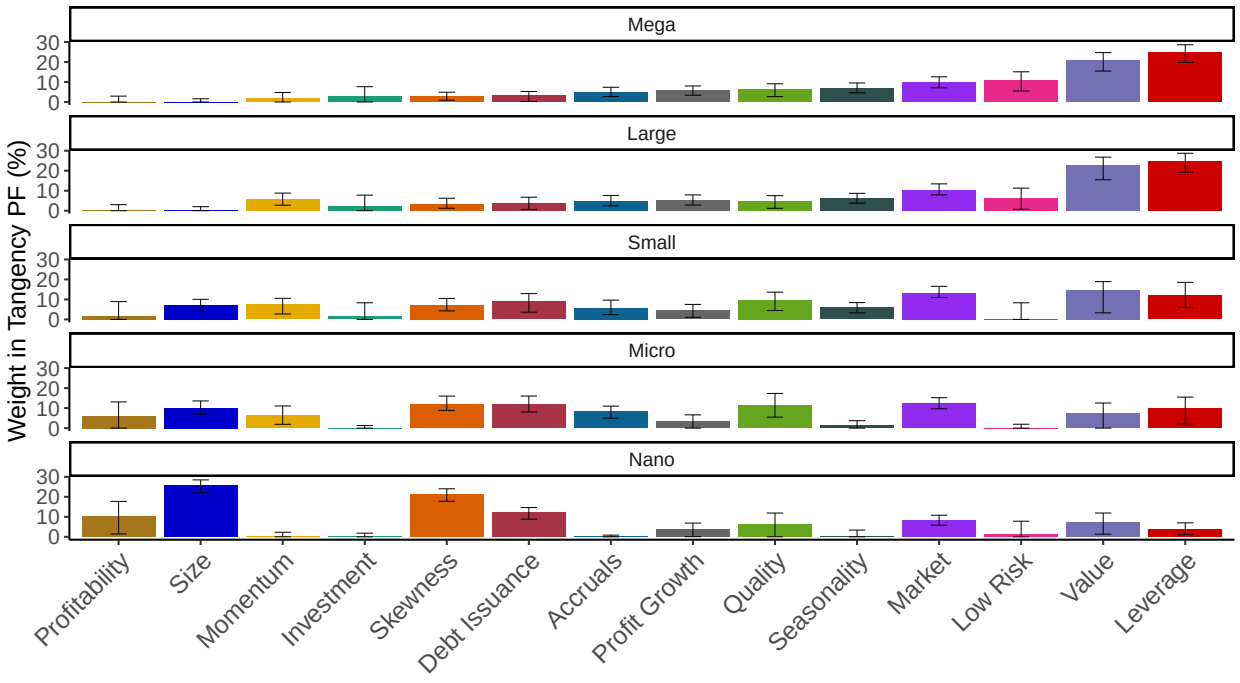

Figure C.11: Tangency Portfolio Weights across Size Groups

Note: Within each size group, we compute the cluster return as the equal weighted return of all factors with data available at a given point in time. We only use US data. We add the US market return. We estimate the tangency weights following the method of Britten-Jones (1999) with a non-negativity constraint. The error bars are the 90% confidence intervals based on 10,000 bootstrap samples and the percentile method. The data starts in 1963.

Table C.2: The Economic Benefit of More Powerful Tests

	Region		
	US	Developed ex. US	Emerging
	(1)	(2)	(3)
Alpha	0.417***	0.282***	0.341***
	[4.209]	[4.283]	[3.426]
Market Beta	−0.140***	−0.126***	−0.034**
	[−3.034]	[−5.000]	[−2.087]
Observations	528	408	362
Adjusted R^2	0.09	0.17	0.01

Note: The dependent variable is an equal weighted portfolio of factors that are significant under empirical Bayes, but not under OLS with the Benjamini-Yekutieli adjustment. A factor is significant under empirical Bayes when the probability of a negative alpha is below 2.5%. A factor is significant under Benjamini-Yekutieli when the adjusted two-sided p-value is below 5% and the OLS alpha estimate is positive. The in-sample estimates are based solely on US data. To avoid lookahead bias, factors are only eligible for inclusion in the portfolio, when the in-sample period of the original paper has ended. Starting in 1959, we update the posterior distribution and the OLS estimates by the end of each year using all of the 153 factors with at least 5 years of data. Based on the in-sample estimates, we invest in the marginally significant factors over the subsequent year. The alpha estimate is in percentages. Standard errors are computed following Newey and West (1987) with 6 lags. The stars indicate *p<0.1; **p<0.05; ***p<0.01.

Table C.3: Factor and Cluster Details

Description	Variable Name	Citation	Orig. Sample	Sign	Orig. Signif.
		Accruals			
Change in current operating working capital	cowc_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Operating accruals	oaccruals_at	Sloan (1996)	1962-1991	-1	1
Percent operating accruals	oaccruals_ni	Hafzalla Lundholm and Van Winkle (2011)	1989-2008	-1	1
Total accruals	taccruals_at	Richardson et al. (2005)	1962-2001	-1	1
Percent total accruals	taccruals_ni	Hafzalla Lundholm and Van Winkle (2011)	1989-2008	-1	1
		Debt Issuance			
Abnormal corporate investment	capex_abn	Titman Wei and Xie (2004)	1973-1996	-1	1
Growth in book debt (3 years)	debt_gr3	Lyandres Sun and Zhang (2008)	1970-2005	-1	1
Change in financial liabilities	fnl_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in noncurrent operating liabilities	ncol_gr1a	Richardson et al. (2005)	1962-2001	-1	0
Change in net financial assets	nfna_gr1a	Richardson et al. (2005)	1962-2001	1	1
Net operating assets	noa_at	Hirshleifer et al. (2004)	1964-2002	-1	1
		Investment			
Liquidity of book assets	aliq_at	Ortiz-Molina and Phillips (2014)	1984-2006	-1	0
Asset Growth	at_gr1	Cooper Gulen and Schill (2008)	1968-2003	-1	1
Change in common equity	be_gr1a	Richardson et al. (2005)	1962-2001	-1	1
CAPEX growth (1 year)	capx_gr1	Xie (2001)	1971-1992	-1	0
CAPEX growth (2 years)	capx_gr2	Anderson and Garcia-Feijoo (2006)	1976-1998	-1	1
CAPEX growth (3 years)	capx_gr3	Anderson and Garcia-Feijoo (2006)	1976-1998	-1	1
Change in current operating assets	coa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in current operating liabilities	col_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Hiring rate	emp_gr1	Belo Lin and Bazdresch (2014)	1965-2010	-1	1
Inventory growth	inv_gr1	Belo and Lin (2011)	1965-2009	-1	1
Inventory change	inv_gr1a	Thomas and Zhang (2002)	1970-1997	-1	1
Change in long-term net operating assets	lnoa_gr1a	Fairfield Whisenant and Yohn (2003)	1964-1993	-1	1
Mispricing factor: Management	mispricing_mgmt	Stambaugh and Yuan (2016)	1967-2013	1	1
Change in noncurrent operating assets	ncoa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in net noncurrent operating assets	nncoa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in net operating assets	noa_gr1a	Hirshleifer et al. (2004)	1964-2002	-1	1
Change PPE and Inventory	ppeinv_gr1a	Lyandres Sun and Zhang (2008)	1970-2005	-1	1
Long-term reversal	ret_60_12	De Bondt and Thaler (1985)	1926-1982	-1	1
Sales Growth (1 year)	sale_gr1	Lakonishok Shleifer and Vishny (1994)	1968-1989	-1	1
Sales Growth (3 years)	sale_gr3	Lakonishok Shleifer and Vishny (1994)	1968-1989	-1	1
Sales growth (1 quarter)	saleq_gr1		1967-2016	-1	0
Years 2-5 lagged returns, nonannual	seas_2_5na	Heston and Sadka (2008)	1965-2002	-1	1

		Leverage			
Firm age	age	Jiang Lee and Zhang (2005)	1965-2001	-1	1
Liquidity of market assets	aliq_mat	Ortiz-Molina and Phillips (2014)	1984-2006	-1	0
Book leverage	at_be	Fama and French (1992)	1963-1990	-1	0
The high-low bid-ask spread	bidaskhl_21d	Corwin and Schultz (2012)	1927-2006	1	1
Cash-to-assets	cash_at	Palazzo (2012)	1972-2009	1	0
Net debt-to-price	netdebt_me	Penman Richardson and Tuna (2007)	1962-2001	-1	1
Earnings volatility	ni_ivol	Francis et al. (2004)	1975-2001	1	0
R&D-to-sales	rd_sale	Chan Lakonishok and Sougiannis (2001)	1975-1995	1	0
R&D capital-to-book assets	rd5_at	Li (2011)	1952-2004	1	0
Asset tangibility	tangibility	Hahn and Lee (2009)	1973-2001	1	0
Altman Z-score	z_score	Dichev (1998)	1981-1995	1	1
		Low Risk			
Market Beta	beta_60m	Fama and MacBeth (1973)	1935-1968	-1	1
Dimson beta	beta_dimson_21d	Dimson (1979)	1955-1974	-1	0
Frazzini-Pedersen market beta	betabab_1260d	Frazzini and Pedersen (2014)	1926-2012	-1	1
Downside beta	betadown_252d	Ang Chen and Xing (2006)	1963-2001	-1	1
Net stock issues	chcsho_12m	Pontiff and Woodgate (2008)	1970-2003	-1	1
Earnings variability	earnings_variability	Francis et al. (2004)	1975-2001	-1	0
Net equity issuance	eqnetis_at	Bradshaw Richardson and Sloan (2006)	1971-2000	-1	1
Free cash flow-to-price	fcf_me	Lakonishok Shleifer and Vishny (1994)	1963-1990	1	1
Idiosyncratic volatility from the CAPM (21 days)	ivol_capm_21d		1967-2016	-1	0
Idiosyncratic volatility from the CAPM (252 days)	ivol_capm_252d	Ali Hwang and Trombley (2003)	1976-1997	-1	1
Idiosyncratic volatility from the Fama-French 3-factor model	ivol_ff3_21d	Ang et al. (2006)	1963-2000	-1	1
Idiosyncratic volatility from the q-factor model	ivol_hxz4_21d		1967-2016	-1	0
Net total issuance	netis_at	Bradshaw Richardson and Sloan (2006)	1971-2000	-1	1
Cash flow volatility	ocfq_saleq_std	Huang (2009)	1980-2004	-1	1
Maximum daily return	rmax1_21d	Bali Cakici and Whitelaw (2011)	1962-2005	-1	1
Highest 5 days of return	rmax5_21d	Bali, Brown, Murray and Tang (2017)	1993-2012	-1	1
Return volatility	rvol_21d	Ang et al. (2006)	1963-2000	-1	1
Share turnover	turnover_126d	Datar Naik and Radcliffe (1998)	1963-1991	-1	1
Number of zero trades with turnover as tiebreaker (6 months)	zero_trades_126d	Liu (2006)	1963-2003	1	1
Number of zero trades with turnover as tiebreaker (1 month)	zero_trades_21d	Liu (2006)	1963-2003	1	0
Number of zero trades with turnover as tiebreaker (12 months)	zero_trades_252d	Liu (2006)	1963-2003	1	1
		Momentum			
Current price to high price over last year	prc_highprc_252d	George and Hwang (2004)	1963-2001	1	1
Residual momentum t-12 to t-1	resff3_12_1	Blitz Huij and Martens (2011)	1930-2009	1	1
Residual momentum t-6 to t-1	resff3_6_1	Blitz Huij and Martens (2011)	1930-2009	1	1

Price momentum t-12 to t-1	ret_12_1	Fama and French (1996)	1963-1993	1	1
Price momentum t-3 to t-1	ret_3_1	Jegedeesh and Titman (1993)	1965-1989	1	1
Price momentum t-6 to t-1	ret_6_1	Jegadeesh and Titman (1993)	1965-1989	1	1
Price momentum t-9 to t-1	ret_9_1	Jegedeesh and Titman (1993)	1965-1989	1	1
Year 1-lagged return, nonannual	seas_1_1na	Heston and Sadka (2008)	1965-2002	1	1
		Profit Growth			
Change sales minus change Inventory	dsale_dinv	Abarbanell and Bushee (1998)	1974-1988	1	1
Change sales minus change receivables	dsale_drec	Abarbanell and Bushee (1998)	1974-1988	-1	0
Change sales minus change SG&A	dsale_dsga	Abarbanell and Bushee (1998)	1974-1988	1	0
Change in quarterly return on assets	niq_at_chg1		1972-2016	1	0
Change in quarterly return on equity	niq_be_chg1		1967-2016	1	0
Standardized earnings surprise	niq_su	Foster Olsen and Shevlin (1984)	1974-1981	1	1
Change in operating cash flow to assets	ocf_at_chg1	Bouchard, Krueger, Landier and Thesmar (2019)	1990-2015	1	1
Quality minus Junk: Safety	qmj_safety	Assness, Frazzini and Pedersen (2018)	1957-2016	1	1
Price momentum t-12 to t-7	ret_12_7	Novy-Marx (2012)	1925-2010	1	1
Labor force efficiency	sale_emp_gr1	Abarbanell and Bushee (1998)	1974-1988	1	0
Standardized Revenue surprise	saleq_su	Jegadeesh and Livnat (2006)	1987-2003	1	1
Year 1-lagged return, annual	seas_1_1an	Heston and Sadka (2008)	1965-2002	1	1
Change in short-term investments	sti_gr1a	Richardson et al. (2005)	1962-2001	1	0
		Profitability			
Coefficient of variation for dollar trading volume	dolvol_var_126d	Chordia Subrahmanyam and Anshuman (2001)	1966-1995	-1	1
Return on net operating assets	ebit_bev	Soliman (2008)	1984-2002	1	1
Profit margin	ebit_sale	Soliman (2008)	1984-2002	1	1
Pitroski F-score	f_score	Piotroski (2000)	1976-1996	1	1
Intrinsic value-to-market	intrinsic_value	Frankel and Lee (1998)	1975-1993	1	0
Return on equity	ni_be	Haugen and Baker (1996)	1979-1993	1	1
Quarterly return on equity	niq_be	Hou Xue and Zhang (2015)	1972-2012	1	1
Ohlson O-score	o_score	Dichev (1998)	1981-1995	-1	1
Operating cash flow to assets	ocf_at	Bouchard, Krueger, Landier and Thesmar (2019)	1990-2015	1	1
Operating profits-to-book equity	ope_be	Fama and French (2015)	1963-2013	1	1
Operating profits-to-lagged book equity	ope_bel1		1967-2016	1	0
Coefficient of variation for share turnover	turnover_var_126d	Chordia Subrahmanyam and Anshuman (2001)	1966-1995	-1	1
		Quality			
Capital turnover	at_turnover	Haugen and Baker (1996)	1979-1993	1	0
Cash-based operating profits-to-book assets	cop_at		1967-2016	1	0
Cash-based operating profits-to-lagged book assets	cop_atl1	Ball et al. (2016)	1963-2014	1	1
Change gross margin minus change sales	dgp_dsale	Abarbanell and Bushee (1998)	1974-1988	1	0
Gross profits-to-assets	gp_at	Novy-Marx (2013)	1963-2010	1	1

Gross profits-to-lagged assets	gp_atl1		1967-2016	1	0
Mispricing factor: Performance	mispricing_perf	Stambaugh and Yuan (2016)	1967-2013	1	1
Number of consecutive quarters with earnings increases	ni_inc8q	Barth Elliott and Finn (1999)	1982-1992	1	0
Quarterly return on assets	niq_at	Balakrishnan Bartov and Faurel (2010)	1976-2005	1	1
Operating profits-to-book assets	op_at		1963-2013	1	1
Operating profits-to-lagged book assets	op_atl1	Ball et al. (2016)	1963-2014	1	1
Operating leverage	opex_at	Novy-Marx (2011)	1963-2008	1	1
Quality minus Junk: Composite	qmj	Assness, Frazzini and Pedersen (2018)	1957-2016	1	1
Quality minus Junk: Growth	qmj_growth	Assness, Frazzini and Pedersen (2018)	1957-2016	1	1
Quality minus Junk: Profitability	qmj_prof	Assness, Frazzini and Pedersen (2018)	1957-2016	1	1
Assets turnover	sale_bev	Soliman (2008)	1984-2002	1	1
Tax expense surprise	tax_gr1a	Thomas and Zhang (2011)	1977-2006	1	1
		Seasonality			
Market correlation	corr_1260d	Assness, Frazzini, Gormsen, Pedersen (2020)	1925-2015	-1	1
Coskewness	coskew_21d	Harvey and Siddique (2000)	1963-1993	-1	1
Net debt issuance	dbnetis_at	Bradshaw Richardson and Sloan (2006)	1971-2000	-1	1
Kaplan-Zingales index	kz_index	Lamont Polk and Saa-Requejo (2001)	1968-1995	1	1
Change in long-term investments	lti_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Earnings persistence	ni_ar1	Francis et al. (2004)	1975-2001	1	0
Taxable income-to-book income	pi_nix	Lev and Nissim (2004)	1973-2000	1	1
Years 11-15 lagged returns, annual	seas_11_15an	Heston and Sadka (2008)	1965-2002	1	1
Years 11-15 lagged returns, nonannual	seas_11_15na	Heston and Sadka (2008)	1965-2002	-1	0
Years 16-20 lagged returns, annual	seas_16_20an	Heston and Sadka (2008)	1965-2002	1	1
Years 16-20 lagged returns, nonannual	seas_16_20na	Heston and Sadka (2008)	1965-2002	-1	1
Years 2-5 lagged returns, annual	seas_2_5an	Heston and Sadka (2008)	1965-2002	1	1
Years 6-10 lagged returns, annual	seas_6_10an	Heston and Sadka (2008)	1965-2002	1	1
Years 6-10 lagged returns, nonannual	seas_6_10na	Heston and Sadka (2008)	1965-2002	-1	1
		Size			
Amihud Measure	ami_126d	Amihud (2002)	1964-1997	1	1
Dollar trading volume	dolvol_126d	Brennan Chordia and Subrahmanyam (1998)	1966-1995	-1	1
Market Equity	market_equity	Banz (1981)	1926-1975	-1	1
Price per share	prc	Miller and Scholes (1982)	1940-1978	-1	1
R&D-to-market	rd_me	Chan Lakonishok and Sougiannis (2001)	1975-1995	1	1
		Skewness			
Idiosyncratic skewness from the CAPM	iskew_capm_21d		1967-2016	-1	0

Idiosyncratic skewness from the Fama-French 3-factor model	iskew_ff3_21d	Bali Engle and Murray (2016)	1925-2021	-1	1
Idiosyncratic skewness from the q-factor model	iskew_hxz4_21d		1967-2016	-1	0
Short-term reversal	ret_1_0	Jegadeesh (1990)	1929-1982	-1	1
Highest 5 days of return scaled by volatility	rmax5_rvol_21d	Assness, Frazzini, Gormsen, Pedersen (2020)	1925-2015	-1	1
Total skewness	rskew_21d	Bali Engle and Murray (2016)	1925-2021	-1	1
		Value			
Assets-to-market	at_me	Fama and French (1992)	1963-1990	1	0
Book-to-market equity	be_me	Rosenberg Reid and Lanstein (1985)	1973-1984	1	1
Book-to-market enterprise value	bev_mev	Penman Richardson and Tuna (2007)	1962-2001	1	1
Debt-to-market	debt_me	Bhandari (1988)	1948-1979	1	1
Dividend yield	div12m_me	Litzenberger and Ramaswamy (1979)	1940-1980	1	1
Ebitda-to-market enterprise value	ebitda_mev	Loughran and Wellman (2011)	1963-2009	1	1
Equity duration	eq_dur	Dechow Sloan and Soliman (2004)	1962-1998	-1	1
Equity net payout	eqnpo_12m	Daniel and Titman (2006)	1968-2003	1	1
Net payout yield	eqnpo_me	Boudoukh et al. (2007)	1984-2003	1	1
Payout yield	eqpo_me	Boudoukh et al. (2007)	1984-2003	1	1
Earnings-to-price	ni_me	Basu (1983)	1963-1979	1	1
Operating cash flow-to-market	ocf_me	Desai Rajgopal and Venkatachalam (2004)	1973-1997	1	1
Sales-to-market	sale_me	Barbee Mukherji and Raines (1996)	1979-1991	1	1

Note: This table shows cluster names as underlined section headings and, for each cluster, a description of the factors included, the variable name used in the code, the original reference, the sample period used in the original reference, the sign of the factor ("1" means "long", "-1" means "short"), and whether the original reference found the factor to be significant ("1" means "yes", "0" means "no"). For example, the first value factor "at_me" goes long stocks with high values of assets-to-market and shorts those with low values (and would be done the reverse if the sign was "-1" instead of "1").

Table C.4: Alpha Across Regions

		US			Developed ex. US			Emerging		
	Factor	α_{OLS}	α_{EB}	$\Pr(\alpha_{\text{EB}} < 0)$	α_{OLS}	α_{EB}	$\Pr(\alpha_{\text{EB}} < 0)$	α_{OLS}	α_{EB}	$\Pr(\alpha_{\text{EB}} < 0)$
1	aliq_mat*	-0.40	-0.31	1.00	-0.38	-0.27	1.00	-0.34	-0.29	1.00
2	bidaskhl_21d	-0.33	-0.29	1.00	-0.67	-0.42	1.00	-0.68	-0.45	1.00
3	dsale_drec*	-0.28	-0.22	1.00	-0.11	-0.13	0.90	-0.15	-0.15	0.93
4	ni_ivol*	-0.25	-0.16	0.98	-0.32	-0.15	0.93	-0.01	-0.09	0.81
5	age	-0.23	-0.15	0.99	-0.23	-0.13	0.91	-0.20	-0.15	0.93
6	at_be*	-0.18	-0.06	0.82	-0.01	0.06	0.25	0.31	0.12	0.12
7	kz_index	-0.11	-0.13	0.94	-0.11	-0.11	0.88	-0.29	-0.15	0.94
8	prc	-0.11	-0.04	0.70	0.05	0.05	0.31	0.12	0.06	0.28
9	turnover_var_126d	-0.10	-0.11	0.95	0.00	-0.02	0.56	0.14	0.00	0.48
10	sti_gr1a*	-0.06	-0.04	0.65	-0.07	-0.01	0.56	0.08	0.02	0.43
11	dolvol_var_126d	-0.05	-0.06	0.83	-0.00	-0.00	0.50	0.15	0.03	0.40
12	netdebt_me	-0.05	0.05	0.23	0.06	0.13	0.08	0.30	0.16	0.05
13	dsale_dsga*	-0.04	0.01	0.45	0.13	0.10	0.16	0.23	0.11	0.14
14	z_score	-0.03	0.05	0.24	-0.08	0.06	0.27	0.20	0.10	0.15
15	ni_ar1*	-0.02	-0.03	0.66	-0.11	-0.04	0.67	-0.03	-0.02	0.59
16	rd_sale*	-0.01	0.07	0.16	0.12	0.15	0.06	0.07	0.09	0.20
17	cash_at*	0.01	0.09	0.11	0.03	0.13	0.09	0.23	0.15	0.07
18	sale_emp_gr1*	0.01	-0.01	0.54	-0.16	-0.04	0.66	0.07	0.01	0.47
19	iskew_hxz4_21d*	0.01	-0.01	0.54	-0.09	0.03	0.41	-0.06	0.02	0.45
20	intrinsic_value*	0.01	-0.03	0.65	0.01	-0.01	0.53	-0.07	-0.05	0.70
21	market_equity	0.02	0.10	0.10	0.12	0.16	0.04	0.51	0.29	0.00
22	ami_126d	0.03	0.10	0.09	0.12	0.15	0.07	0.38	0.22	0.02
23	ncol_gr1a*	0.05	-0.02	0.62	-0.06	0.02	0.41	0.05	0.06	0.28
24	iskew_ff3_21d	0.10	0.06	0.25	-0.17	0.01	0.46	0.21	0.10	0.20
25	rd5_at*	0.11	0.20	0.00	0.29	0.30	0.00	0.44	0.28	0.01
26	coskew_21d	0.11	0.10	0.10	0.29	0.16	0.05	-0.02	0.08	0.21
27	zero_trades_21d*	0.11	0.03	0.33	0.38	0.06	0.31	-0.28	-0.06	0.70
28	lti_gr1a	0.11	0.06	0.23	0.02	0.03	0.38	-0.10	0.01	0.48
29	ni_inc8q*	0.12	0.18	0.02	0.45	0.30	0.01	0.33	0.24	0.02
30	tax_gr1a	0.12	0.15	0.03	0.03	0.13	0.10	0.36	0.20	0.03
31	seas_16_20na	0.13	0.10	0.12	-0.06	0.08	0.25	0.34	0.13	0.15
32	tangibility*	0.13	0.22	0.00	0.26	0.30	0.00	0.41	0.30	0.00
33	ret_60_12	0.13	0.01	0.45	0.20	0.17	0.05	0.31	0.23	0.01
34	debt_me	0.13	0.04	0.29	0.08	0.01	0.46	-0.15	-0.09	0.80
35	saleq_gr1*	0.14	-0.05	0.73	0.07	0.01	0.46	-0.48	-0.10	0.80
36	col_gr1a	0.15	-0.01	0.57	0.07	0.05	0.29	-0.10	0.03	0.38
37	gp_atl1*	0.15	0.20	0.01	0.19	0.24	0.01	0.49	0.30	0.00
38	pi_nix	0.20	0.15	0.03	0.12	0.12	0.10	0.03	0.10	0.15
39	ret_3_1	0.21	0.14	0.03	0.27	0.18	0.03	0.16	0.18	0.04
40	bev_mev	0.21	0.14	0.02	0.29	0.20	0.01	0.27	0.18	0.04
41	opex_at	0.22	0.22	0.00	0.18	0.21	0.02	0.18	0.17	0.04
42	at_me*	0.23	0.15	0.02	0.27	0.18	0.02	0.21	0.14	0.08
43	ebit_sale	0.23	0.20	0.00	0.33	0.27	0.00	0.33	0.24	0.01
44	at_turnover*	0.23	0.25	0.00	0.26	0.28	0.00	0.42	0.29	0.00
45	op_atl1	0.25	0.27	0.00	0.24	0.28	0.00	0.48	0.33	0.00
46	earnings_variability*	0.25	0.16	0.02	0.13	0.09	0.18	0.07	0.10	0.17
47	seas_11_15na*	0.25	0.16	0.02	-0.17	0.05	0.33	-0.00	0.09	0.22
48	dolvol_126d	0.25	0.30	0.00	0.27	0.31	0.00	0.55	0.38	0.00
49	seas_1_1na	0.26	0.18	0.01	0.20	0.14	0.06	0.20	0.20	0.03
50	saleq_su	0.28	0.25	0.00	0.27	0.26	0.01	0.33	0.26	0.01

51	be_me	0.28	0.20	0.00	0.33	0.25	0.00	0.31	0.22	0.02
52	be_gr1a	0.28	0.10	0.09	0.16	0.14	0.08	-0.10	0.09	0.20
53	ope_bel1*	0.29	0.25	0.00	0.23	0.25	0.00	0.48	0.31	0.00
54	div12m_me	0.29	0.25	0.00	0.63	0.48	0.00	0.77	0.51	0.00
55	dbnetis_at	0.29	0.25	0.00	0.14	0.21	0.02	0.45	0.29	0.00
56	beta_dimson_21d*	0.30	0.22	0.00	0.31	0.20	0.02	-0.02	0.11	0.14
57	niq_su	0.30	0.25	0.00	0.13	0.22	0.02	0.38	0.26	0.01
58	sale_gr3	0.30	0.13	0.05	0.15	0.16	0.06	0.12	0.19	0.04
59	corr_1260d	0.30	0.25	0.00	0.22	0.22	0.01	0.20	0.22	0.01
60	rd_me	0.31	0.35	0.00	0.36	0.36	0.00	0.48	0.38	0.00
61	ret_6_1	0.31	0.24	0.00	0.32	0.25	0.00	0.39	0.33	0.00
62	ocfq_saleq_std	0.31	0.26	0.00	0.39	0.30	0.00	0.82	0.40	0.00
63	sale_gr1	0.31	0.13	0.05	0.21	0.16	0.05	-0.18	0.07	0.25
64	dgp_dsale*	0.32	0.26	0.00	-0.09	0.09	0.19	0.08	0.12	0.13
65	seas_2_5na	0.32	0.19	0.00	0.47	0.37	0.00	0.44	0.39	0.00
66	ivol_capm_252d	0.33	0.27	0.00	0.47	0.32	0.00	0.23	0.25	0.01
67	sale_me	0.33	0.25	0.00	0.34	0.27	0.00	0.39	0.27	0.00
68	niq_at	0.33	0.37	0.00	0.25	0.38	0.00	0.91	0.50	0.00
69	qmj_safety	0.34	0.34	0.00	0.31	0.36	0.00	0.73	0.46	0.00
70	ni_be	0.34	0.29	0.00	0.40	0.33	0.00	0.27	0.25	0.01
71	gp_at	0.34	0.37	0.00	0.30	0.38	0.00	0.78	0.50	0.00
72	aliq_at*	0.35	0.16	0.02	0.08	0.12	0.11	-0.02	0.13	0.11
73	o_score	0.35	0.31	0.00	0.36	0.34	0.00	0.54	0.38	0.00
74	prc_highprc_252d	0.35	0.28	0.00	0.37	0.29	0.00	0.36	0.33	0.00
75	zero_trades_126d	0.36	0.27	0.00				-0.00	0.16	0.08
76	turnover_126d	0.36	0.31	0.00	0.53	0.37	0.00	0.46	0.36	0.00
77	taccruals_at	0.36	0.17	0.02	0.07	0.14	0.08	0.04	0.12	0.12
78	emp_gr1	0.36	0.22	0.00	0.32	0.32	0.00	0.57	0.41	0.00
79	betadown_252d	0.37	0.31	0.00	0.51	0.36	0.00	0.34	0.31	0.00
80	ret_1_0	0.37	0.31	0.00	0.23	0.29	0.00	0.18	0.25	0.01
81	beta_60m	0.37	0.32	0.00	0.42	0.34	0.00	0.56	0.39	0.00
82	sale_bev	0.38	0.37	0.00	0.31	0.35	0.00	0.41	0.34	0.00
83	seas_2_5an	0.39	0.37	0.00	0.39	0.37	0.00	0.56	0.42	0.00
84	seas_16_20an	0.40	0.33	0.00	0.34	0.30	0.00	0.20	0.26	0.00
85	eq_dur	0.40	0.32	0.00	0.45	0.36	0.00	0.49	0.36	0.00
86	niq_at_chg1*	0.40	0.40	0.00	0.54	0.45	0.00	0.67	0.46	0.00
87	ret_9_1	0.41	0.34	0.00	0.44	0.35	0.00	0.43	0.39	0.00
88	iskew_capm_21d*	0.41	0.33	0.00	-0.08	0.15	0.06	0.29	0.26	0.00
89	seas_6_10an	0.41	0.38	0.00	0.80	0.50	0.00	0.11	0.30	0.00
90	seas_11_15an	0.41	0.30	0.00	0.30	0.23	0.01	-0.16	0.11	0.13
91	ope_be	0.43	0.38	0.00	0.43	0.39	0.00	0.48	0.38	0.00
92	betabab_1260d	0.43	0.38	0.00	0.63	0.47	0.00	0.65	0.48	0.00
93	seas_1_1an	0.43	0.38	0.00	0.43	0.37	0.00	0.27	0.32	0.00
94	taccruals_ni	0.44	0.20	0.01	-0.02	0.08	0.20	-0.17	0.03	0.38
95	zero_trades_252d	0.44	0.35	0.00				-0.00	0.20	0.04
96	op_at	0.44	0.46	0.00	0.53	0.50	0.00	0.58	0.48	0.00
97	netis_at	0.44	0.38	0.00	0.49	0.41	0.00	0.78	0.50	0.00
98	ivol_capm_21d*	0.45	0.39	0.00	0.58	0.44	0.00	0.53	0.44	0.00
99	niq_be	0.45	0.39	0.00	0.60	0.44	0.00	0.33	0.34	0.00
100	ebit_bev	0.45	0.40	0.00	0.33	0.36	0.00	0.63	0.44	0.00
101	qmj_prof	0.46	0.45	0.00	0.44	0.45	0.00	0.56	0.45	0.00
102	at_gr1	0.46	0.27	0.00	0.20	0.23	0.01	0.13	0.25	0.01
103	eqpo_me	0.47	0.36	0.00	0.27	0.26	0.00	0.53	0.36	0.00
104	capx_gr3	0.48	0.30	0.00	0.59	0.42	0.00	0.00	0.27	0.01

105	eqnpo_12m	0.48	0.43	0.00	0.79	0.62	0.00	0.71	0.56	0.00
106	ni_me	0.48	0.41	0.00	0.47	0.41	0.00	0.73	0.51	0.00
107	rvol_21d	0.49	0.41	0.00	0.55	0.41	0.00	0.32	0.35	0.00
108	ivol_hxz4_21d*	0.49	0.44	0.00	0.75	0.53	0.00	1.06	0.57	0.00
109	qmj_growth	0.49	0.43	0.00	0.22	0.30	0.00	0.14	0.25	0.01
110	niq_be_chg1*	0.49	0.44	0.00	0.23	0.38	0.00	0.84	0.48	0.00
111	ret_12_7	0.49	0.46	0.00	0.55	0.48	0.00	0.51	0.46	0.00
112	ivol_ff3_21d	0.50	0.44	0.00	0.52	0.45	0.00	0.93	0.54	0.00
113	seas_6_10na	0.50	0.43	0.00	0.69	0.46	0.00	0.13	0.34	0.00
114	ebitda_mev	0.50	0.42	0.00	0.47	0.41	0.00	0.72	0.51	0.00
115	capex_abn	0.51	0.36	0.00	0.31	0.35	0.00	0.18	0.32	0.00
116	ret_12_1	0.52	0.43	0.00	0.45	0.38	0.00	0.42	0.41	0.00
117	qmj	0.53	0.51	0.00	0.55	0.51	0.00	0.46	0.45	0.00
118	eqnetis_at	0.54	0.45	0.00	0.60	0.47	0.00	0.60	0.48	0.00
119	ocf_me	0.54	0.46	0.00	0.41	0.39	0.00	0.85	0.57	0.00
120	coa_gr1a	0.54	0.35	0.00	0.25	0.30	0.00	0.29	0.35	0.00
121	rskew_21d	0.54	0.44	0.00	0.08	0.25	0.00	0.13	0.26	0.00
122	ocf_at_chg1	0.55	0.48	0.00	0.61	0.50	0.00	0.31	0.40	0.00
123	mispricing_perf	0.56	0.54	0.00	0.52	0.51	0.00	0.51	0.47	0.00
124	chcsho_12m	0.57	0.49	0.00	0.70	0.53	0.00	0.44	0.45	0.00
125	f_score	0.57	0.50	0.00	0.50	0.48	0.00	0.66	0.51	0.00
126	eqnpo_me	0.58	0.47	0.00	0.43	0.39	0.00	0.72	0.50	0.00
127	rmax1_21d	0.58	0.50	0.00	0.62	0.50	0.00	0.52	0.48	0.00
128	rmax5_21d	0.59	0.52	0.00	0.67	0.52	0.00	0.42	0.44	0.00
129	resff3_6_1	0.61	0.51	0.00	0.67	0.55	0.00	0.64	0.56	0.00
130	capx_gr2	0.63	0.43	0.00	0.51	0.45	0.00	0.25	0.40	0.00
131	fcf_me	0.63	0.57	0.00	0.61	0.56	0.00	1.10	0.71	0.00
132	capx_gr1*	0.65	0.45	0.00	0.39	0.40	0.00	0.31	0.42	0.00
133	rmax5_rvol_21d	0.69	0.58	0.00	0.34	0.43	0.00	0.10	0.34	0.00
134	oaccruals_at	0.71	0.52	0.00	0.42	0.50	0.00	0.71	0.59	0.00
135	debt_gr3	0.71	0.55	0.00	0.54	0.53	0.00	0.35	0.49	0.00
136	ppeinv_gr1a	0.71	0.50	0.00	0.42	0.43	0.00	0.25	0.41	0.00
137	fnl_gr1a	0.73	0.55	0.00	0.29	0.43	0.00	0.45	0.49	0.00
138	lnoa_gr1a	0.74	0.52	0.00	0.46	0.45	0.00	0.18	0.39	0.00
139	cop_atl1	0.75	0.69	0.00	0.47	0.56	0.00	0.69	0.60	0.00
140	inv_gr1a	0.75	0.53	0.00	0.36	0.42	0.00	0.39	0.47	0.00
141	nfna_gr1a	0.77	0.58	0.00	0.23	0.43	0.00	0.57	0.54	0.00
142	ocf_at	0.78	0.72	0.00	0.90	0.78	0.00	0.88	0.74	0.00
143	oaccruals_ni	0.78	0.55	0.00	0.18	0.38	0.00	0.60	0.52	0.00
144	dsale_dinv	0.79	0.59	0.00	0.34	0.40	0.00	0.01	0.31	0.00
145	ncoa_gr1a	0.82	0.58	0.00	0.32	0.41	0.00	0.31	0.45	0.00
146	mispricing_mgmt	0.83	0.63	0.00	0.67	0.63	0.00	0.51	0.60	0.00
147	nncoa_gr1a	0.84	0.59	0.00	0.27	0.38	0.00	0.28	0.43	0.00
148	inv_gr1	0.84	0.61	0.00	0.52	0.50	0.00	0.09	0.40	0.00
149	noa_at	0.85	0.61	0.00	0.36	0.45	0.00	0.09	0.39	0.00
150	cowc_gr1a	0.85	0.61	0.00	0.36	0.47	0.00	0.49	0.52	0.00
151	resff3_12_1	0.94	0.83	0.00	1.15	0.94	0.00	0.65	0.78	0.00
152	noa_gr1a	0.96	0.73	0.00	0.60	0.62	0.00	0.50	0.62	0.00
153	cop_at*	1.04	0.94	0.00	0.68	0.75	0.00	0.79	0.76	0.00

Note: The table shows monthly alpha in percentages across three different regions. α_{OLS} is the intercept from an OLS regression of the factor return on the regional market return. α_{EB} is the factor-region specific posterior mean found via the empricial Bayes procedure applied jointly to all the factor-region specific factors. $\Pr(\alpha_{\text{EB}} < 0)$ is the probability that the alpha is negative based on the posterior distribution from the EB procedure. We count a factor as replicated if this probability is below 2.5%. The residual volatility of all strategies have been scaled to 10% annualized. A "*" indicates that the original paper did not propose the factors as a significant predictor of realized returns.

Table C.5: Country Information

	Country	MSCI	Start	Stocks	Mega Stocks	Total Market Cap	Median MC
1	USA	Developed	1926-01-31	5,256	414	3.51e+07	407
2	CHN	Emerging	1991-02-28	3,665	67	7.13e+06	699
3	JPN	Developed	1986-01-31	3,831	85	6.35e+06	180
4	HKG	Developed	1986-01-31	2,269	52	4.52e+06	104
5	GBR	Developed	1986-01-31	1,702	35	3.14e+06	138
6	FRA	Developed	1986-01-31	693	36	2.77e+06	108
7	DEU	Developed	1986-01-31	674	33	2.35e+06	118
8	IND	Emerging	1988-09-30	3,397	28	2.18e+06	8
9	CAN	Developed	1982-03-31	714	36	2.05e+06	156
10	CHE	Developed	1986-01-31	236	21	1.52e+06	794
11	AUS	Developed	1985-11-30	1,669	17	1.46e+06	32
12	KOR	Emerging	1986-02-28	2,185	17	1.43e+06	96
13	TWN	Emerging	1988-02-29	1,928	12	1.34e+06	98
14	NLD	Developed	1986-01-31	119	16	1.01e+06	931
15	RUS	Emerging	1995-08-31	199	10	7.59e+05	171
16	SWE	Developed	1986-01-31	652	14	7.53e+05	65
17	ITA	Developed	1986-01-31	357	11	7.40e+05	134
18	ESP	Developed	1986-01-31	183	14	7.35e+05	301
19	BRA	Emerging	1988-05-31	193	8	6.43e+05	825
20	SGP	Developed	1986-01-31	570	8	5.80e+05	56
21	THA	Emerging	1986-07-31	740	6	5.49e+05	76
22	SAU	Emerging	2000-02-29	193	8	5.25e+05	307
23	IDN	Emerging	1989-01-31	626	7	5.16e+05	101
24	ZAF	Emerging	1986-01-31	276	5	4.46e+05	162
25	DNK	Developed	1986-01-31	155	6	4.22e+05	131
26	MYS	Emerging	1986-01-31	911	3	4.11e+05	41
27	BEL	Developed	1986-01-31	129	5	3.98e+05	441
28	MEX	Emerging	1986-02-28	113	5	3.38e+05	1,040
29	NOR	Developed	1986-01-31	247	3	3.22e+05	198
30	PHL	Emerging	1986-01-31	246	2	2.68e+05	140
31	FIN	Developed	1986-01-31	153	6	2.53e+05	154
32	ARE	Emerging	2001-06-30	103	4	2.35e+05	295
33	ISR	Developed	1994-12-31	408	0	2.05e+05	81
34	CHL	Emerging	1989-01-31	174	1	1.88e+05	237
35	TUR	Emerging	1990-03-31	395	1	1.86e+05	61
36	QAT	Emerging	2001-12-31	46	2	1.60e+05	957
37	POL	Emerging	1993-07-31	689	0	1.54e+05	9
38	VNM	Frontier	2006-08-31	660	1	1.49e+05	15
39	IRL	Developed	1986-01-31	34	3	1.30e+05	553
40	AUT	Developed	1986-01-31	62	2	1.28e+05	538
41	COL	Emerging	1989-01-31	45	1	1.08e+05	554
42	NZL	Developed	1986-01-31	123	0	1.07e+05	192
43	KWT	Frontier	2001-04-30	164	2	1.07e+05	78
44	PER	Emerging	1990-01-31	101	2	1.07e+05	81
45	ARG	Emerging	1988-09-30	69	1	7.31e+04	95
46	PRT	Developed	1986-08-31	42	1	7.09e+04	126
47	MAR	Frontier	1995-09-30	68	0	6.42e+04	164
48	GRC	Emerging	1988-09-30	154	0	5.50e+04	29
49	PAK	Emerging	1992-09-30	419	0	5.03e+04	15
50	EGY	Emerging	1996-12-31	201	0	4.22e+04	33
51	NGA	Frontier	1993-11-30	152	0	3.32e+04	9

52	BGD	Frontier	2002-05-31	318	0	3.28e+04	23
53	HUN	Emerging	1993-06-30	35	0	3.20e+04	53
54	CZE	Emerging	1995-01-31	10	0	2.72e+04	1,216
55	ROU	Frontier	1997-11-30	70	0	2.56e+04	39
56	KEN	Frontier	1993-11-30	49	0	2.46e+04	54
57	BHR	Frontier	2001-03-31	24	0	2.18e+04	271
58	HRV	Frontier	1997-11-30	65	0	2.02e+04	62
59	JOR	Frontier	1993-08-31	155	0	1.97e+04	17
60	TTO	Standalone	1997-08-31	20	0	1.75e+04	473
61	OMN	Frontier	1998-03-31	95	0	1.67e+04	44
62	LKA	Frontier	1987-07-31	262	0	1.42e+04	12
63	KAZ	Frontier	2009-06-30	11	0	1.19e+04	758
64	ISL	Standalone	1995-12-31	21	0	1.02e+04	235
65	JAM	Standalone	1993-12-31	38	0	1.02e+04	77
66	MUS	Frontier	1995-08-31	60	0	9.16e+03	70
67	TUN	Frontier	1995-09-30	71	0	8.39e+03	32
68	LUX	Not Rated	1986-01-31	11	0	7.85e+03	367
69	SVN	Frontier	1995-03-31	19	0	7.71e+03	130
70	CIV	Frontier	2002-05-31	38	0	7.37e+03	81
71	MLT	Standalone	1995-08-31	20	0	5.25e+03	175
72	BGR	Standalone	1995-12-31	93	0	4.94e+03	25
73	LTU	Frontier	1995-11-30	26	0	4.07e+03	92
74	BWA	Standalone	1995-09-30	21	0	3.43e+03	99
75	GHA	Not Rated	1997-11-30	15	0	3.40e+03	90
76	PSE	Standalone	2008-08-31	24	0	3.34e+03	100
77	TZA	Not Rated	2000-07-31	10	0	3.19e+03	132
78	EST	Frontier	1996-01-31	18	0	3.02e+03	84
79	NAM	Not Rated	1996-06-30	7	0	2.98e+03	530
80	CYP	Not Rated	1994-01-31	33	0	2.78e+03	26
81	BMU	Not Rated	2009-11-30	5	0	2.74e+03	181
82	SRB	Frontier	2009-09-30	18	0	2.60e+03	25
83	SVK	Not Rated	1986-01-31	9	0	2.58e+03	19
84	ECU	Not Rated	2000-09-30	4	0	2.29e+03	337
85	UKR	Standalone	2008-03-31	15	0	2.00e+03	63
86	LBN	Standalone	1997-11-30	3	0	1.51e+03	566
87	MWI	Not Rated	2008-08-31	8	0	1.37e+03	148
88	VEN	Not Rated	1989-01-31	24	0	1.28e+03	21
89	UGA	Not Rated	2011-10-31	9	0	1.12e+03	101
90	LVA	Not Rated	1997-11-30	14	0	9.17e+02	14
91	ZMB	Not Rated	1996-03-31	11	0	8.67e+02	32
92	GGY	Not Rated	2016-12-31	1	0	1.18e+02	118
93	ZWE	Standalone	1995-08-31	50	0	8.04e+01	1
	All			40,200	1,011	8.37e+07	

Note: The table shows summary statistics by the country where a security is listed. We only include countries with data available by December 31st 2019. We include common stocks that are the primary security of the underlying firm, with non-missing return and lagged market equity data. *Country* is the ISO code of the underlying exchange country. For further information, see https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes. *MSCI* shows the MSCI classification of each country as of January 7th 2021. For the most recent classification, see https://www.msci.com/market-classification. *Start* is the first date with a valid observation. In the next 4 columns, the data is shown as of December 31st 2019. *Stocks* is the number of stocks available. *Mega stocks* is the number of stocks with a market cap above the 80th percentile of NYSE stocks. *Total Market Cap* is the aggregate market cap in million USD. *Median MC* is the median market cap in million USD.

D Identifier Variables

This section covers all of the variables that give firm/date level identifiers and information. If a variable starts with 'comp' or 'crsp', then the following variable name is drawn from the specified dataset. For example, 'crsp_shrcd' is the 'shrcd' variable from CRSP.

Table D.6: Identifier Variables

Name	Description
size_grp	This groups each firm into one of five categories: Mega, Large, Small, Micro and Nano cap. The groups are non-overlapping and the breakpoints are based on the market equity of NYSE stocks. In particular, Mega caps are all stocks with market equity larger than the 80th percentile of NYSE stocks, Large caps are all remaining stocks larger than the 50th percentile, Small caps are larger than the 20th percentile, Micro caps are larger than the 1st percentile and Nano caps are the remaining stocks.
id	Dataset's unique firm identifier variable. It first identifies the source of the data 'crsp' or 'comp' and also a number as a firm identifier.
source	Identifies the source of the firm/date observation which is either CRSP or Compustat
obs_main	If there are more than one firm observations for one date, this identifies if the observation is considered as the 'main' observation. If available, CRSP observations are considered as the 'main' observation.
gvkey	Permanent six-digit unique firm identifier from Compustat
iid	Permanent two-digit addition to 'gvkey' that identifies specific issues of a firm from Compustat
primary_sec	Primary security as identified by Compustat. A 'gvkey' can have up to three different primary securities ('iid)' at a given time (US, CA, and international).
permno	Permanent unique firm identifier from CRSP
permco	Permanent issue identifier from CRSP
excntry	Stock exchange country code from CRSP
curcd	ISO currency code
fx	Ratio of firm currency to USD at the date of observation
common	If CRSP is the source, common is one if the SHRCD variable is 10, 11 or 12. If Compustat is the source, common is one if TPCI is '0'
comp_tpci	Compustat issue type identifier
crsp_shrcd	CRSP share code
comp_exchg	Compustat stock exchange code
crsp_exchg	CRSP stock exchange code
crsp_sic	CRSP firm industry identifier (sic2)
date	Date of the observation
eom	The last day of the month in which the observation is made
adjfct	Share adjustment factor, using 'cfacshr' if the source is CRSP or 'ajexdi' if the source is Compustat

E Variable Definitions and Data Construction

E.1 Helper Functions

This section describes functions that we use to create variables. Many of the functions are used for variables with quarterly, monthly and daily frequencies, and these are specified by "_zQ", "_zM" and "_zD" respectively, where "z" is the number of quarters, months or days that the function is referencing. For example, MEAN_12M(X) is the mean of the past 12 months of variable X.

Table E.7: Helper Functions

Function	Name	Description
Mean	MEAN_z(X)	$\frac{1}{z}\sum_{n=0}^{z-1} X_{t-n}$

Function	Name	Description
Variance	VARC_z(X)	$\frac{1}{z-1}\sum_{n=0}^{z-1}(X_{t-n} - MEAN\_z(X_t))^2$
Covariance	COVAR_z(X, Y)	$\frac{1}{z-1}\sum_{n=0}^{z-1}(X_{t-n} - MEAN\_z(X_t))(Y_{t-n} - MEAN\_z(Y_t))$
Standard Deviation	SDEV_z(X)	$\sqrt{VARC\_z(X)}$
Skewness	SKEW_z(X)	$\frac{1}{z \bullet SDEV\_z(X)^3}\sum_{n=0}^{z-1}(X_{t-n} - MEAN\_z(X_t))^3$
Standardized Unexpected Realization	SUR_z(X)	$\frac{X_t - (X_{t-3} + MEAN\_z(X_{t-3} - X_{t-15})/4)}{SDEV\_z(X_{t-3} - X_{t-15})}$
Change to Expectations	CHG_TO_EXP(X)	$\frac{X_t}{(X_{t-12} + X_{t-24})/2}$
Maximum	MAXn_z(X)	The maximum n values of given input.
Quality Minus Junk Variables		
Earnings Volatility	$\_EVOL$	$ROEQ\_BE\_STD \bullet 2$. If this is unavailable, we use ROE_BE_STD.
Rank of Variable	$\_rVar$	Cross-sectional rank of Var within a country[45]
Z transformation	$ZV(rVar)$	$\frac{\_rVAR - MEAN_t(\_rVAR)}{SDEV\_t(\_rVAR)}$

E.2 Accounting Characteristics

E.2.1 Datasets

- COMP.FUNDA
- COMP.FUNDQ
- COMP.G_FUNDA
- COMP.G_FUNDQ

E.2.2 General Information

- We create characteristics for annual and quarterly accounting data separately. We then take the most recent characteristics value from each dataset to create the final dataset.

[45] $OACCRUALS\_AT$, $BETABAB\_1260d$, $DEBT\_AT$ and $\_EVOL$ are sorted in descending order. All other variables are sorted in ascending order.

- We assume that accounting variables are publically available 4 months after the end of the accounting period .

- In describing accounting variables, we use the Compustat item names from the annual dataset. The equivalent item name in the quarterly dataset can be found by adding a 'q' or 'y' to the end of the annual item name. Specifically, 'q' indicates a value calculated over one quarter while 'y' refers to the cummulative value over the quarters with data available within a fiscal year.

E.2.3 Annualized Accounting Variables from Quarterly Data

- The value of a balance sheet item such as asset or book equity has the same meaning in the annual and the quarterly data. It is the value by the end of a fiscal period.

- The value of an income or cash flow statement item is different. In the annual data, it is calculated over one year. However, in the quarterly data, it is calculated over one quarter. To make quarterly income and cash flows items comparable to the corresponding annual item, we take the sum of the item over the last four quarters.

E.2.4 Accounting Variables

The abbreviation is used to refer to the accounting variable. A suffix of '*' indicates that we have altered the original Compustat item to increase the coverage or to create a variable that is a part of creating a characteristic in the final dataset. The characteristic name will reflect the accounting name except the '*' suffix. As an example, 'gp_at' is gross profit scaled by assets. In general, we will refer to Compustat variables using capital letters.

Table E.8: Accounting Variables

Name	Abbreviation	Construction
Income Statement		
Sales	sale*	We prefer SALE. If this is unavailable, we use REVT
Gross Profit	gp*	We prefer to use GP. If this is unavailable we use sale*-COGS
Selling, General and Administrative Expenses	xsga	Compustat item XSGA
Research and Development Expenses	xrd	Compustat item XRD. Note that this is not available in Compustat Global
Operating Expenses	opex*	We prefer to use XOPR. If this is unavailable, we use COGS+XSGA
Operating Income Before Depreciation	ebitda*	We prefer to use EBITDA. If this is unavailable, we use OIBDP. If this is unavailable, we use SALE*-OPEX*. If this is unavailable, we use GP*-XSGA
Operating Income After Depreciation	ebit*	We prefer to use EBIT. If this is unavailable, we use OIADP. If this is unavailable, we use EBITDA*-DP
Operating Profit ala Ball et al (2015)	op*	We use EBITDA* + XRD. If XRD is unavailable, we set it to zero
Operating Profit to Equity	ope*	We use EBITDA*-XINT. Note that we target the same variable as the numerator of the profitability characteristic used to create the Robust-minus weak factor in the fama-French 5 factor model (Fama and French, 2015)
Earnings before Tax and Extraordinary Items	pi*	We prefer to use PI. If this is unavailable we use EBIT*-XINT+SPI+NOPI where we set SPI and NOPI to zero if missing
Income Tax	tax	Compustat item TXT
Extraordinary Items and Discontinued Operations	xido*	We prefer to use XIDO. If this is unavailable, we use XI+DO where we set DO to zero if missing. The reason why we set missing DO to zero is because it is not available in COMP.G_FUNDQ
Net Income	ni*	We prefer to use IB. If this is unavailable, we use NI-XIDO*. If this is unavailable, we prefer PI*-TXT-MII. If MII is unavailble, it is set to zero
Net Income Including Extraordinary Items	nix*	We prefer NI. If this is not available, we prefer NI*+XIDO*. If XIDO* is unavailable, we set it to zero. If that is unavailable, we prefer NI*+XI+DO
Total Dividends	div*	We prefer DVT. If this is not available, we use DV

Name	Abbreviation	Construction
Income Before Extraordinary Items	ni_qtr*	We use IBQ
Net Sales	sale_qtr*	We use SALEQ
Cash Flow Statement		
Capital Expenditures	capx	Compustat item CAPX
Capital Expenditures to Sales	capx_sale*	We use CAPX / SALE*
Free Cash Flow	fcf*	We use OCF*-CAPX. Note that the free cash flow is computed before financing activities and sale of assets is taken into account
Equity Buyback	eqbb*	We use PRSTKC+PURTSHR Equity Buyback is mainly PRSTKC in NA and PURTSHR in GLOBAL. Either of PRSTKC or PURTSHR are allowed to be missing
Equity Issuance	eqis*	Compustat item SSTK
Equity Net Issuance	eqnetis*	We use EQIS*-EQBB*. Either EQIS* or EQBB* are allowed to be missing
Net Equity Payout	eqpo*	We use DIV*+EQBB*
Equity Net Payout	eqnpo*	We use DIV*-EQNETIS*
Net Long-Term Debt Issuance	dltnetis*	We prefer to use DLTIS-DLTR where we only require that one of the items are non-missing. If this is unavailable, we use LTDCH. If this is unavailable we use the yearly change in long-term book debt DLTT
Net Short-Term Debt Issuance	dstnetis*	We prefer DLCCH. If this is unavailable, we use the yearly change in short-term book debt DLC
Net Debt Issuance	dbnetis*	We use DLTNETIS*+DSTNETIS* and only require one of the items to be non-missing
Net Issuance	netis*	We use EQNETIS*+DBNETIS*. Either EQNETIS* or DBNETIS* are allowed to be missing
Balance Sheet - Assets		
Total Assets	at*	We prefer to use AT. If this is unavailable, then we use SEQ* + DLTT + LCT + LO + TXDITC. If LCT, LO, or TXDITC are missing, then they are set to zero
Current Assets	ca*	We prefer ACT. If this is unavailable, we use RECT+INVT+CHE+ACO
Account Receivables	rec	Compustat item RECT
Cash and Short-Term Investment	cash	Compustat item CHE
Inventory	inv	Compustat item INVT
Investment and Advances	ivao	Compustat item IVAO
Property, Plans and Equipment Gross	ppeg	Compustat item PPEGT
Balance Sheet - Liabilities		
Current Liabilities	cl*	We prefer LCT. If this is unavailable, we use AP+ DLC+ TXP+ LCO
Accounts Payable	ap	Compustat item AP
Deferred Taxes and Investment Credit	txditc*	We prefer to use TXDITC. If this is unavailable, we use TXDB+ ITCB
Balance Sheet - Financing		
Preferred Stock	pstk*	We prefer to use PSTKRV. If this is unavailable, we use PSTKL. If this is unavailable, we use PSTK
Total Debt	debt*	We use DLTT+ DLC. Either DLTT or DLC are allowed to me missing
Net Debt	netdebt*	We use DEBT*- CHE where we set CHE to zero if missing
Shareholders Equity	seq*	We prefer to use SEQ. If this is unavailable, we use CEQ+PSTK* where we set PSTK* to zero if missing. If this is unavailable, we use AT- LT
Book Equity	be*	We use SEQ*+TXDITC*-PSTK* where we set TXDITC* and PSTK* to zero if missing
Book Enterprise Value	bev*	We prefer to use ICAPT+DLC-CHE where DLC and CHE are set to zero if missing. If this is unavailable, we use SEQ*+NETDEBT*+ MIB where we set MIB to zero if missing. In the global data ICAPT is reduced by Treasury stock
Balance Sheet - Summary		
Current Operating Assets	coa*	We use CA*- CHE
Current Operating Liabilities	col*	We use CL*- DLC. If DLC is missing, it is set to zero
Current Operating Working Capital	cowc*	We use COA*-COL*
Non-Current Operating Assets	ncoa*	We use AT* - CA*- IVAO
Non-Current Operating Liabilities	ncol*	We use LT-CL*- DLTT
Net Non-Current Operating Assets	nncoa*	We use NCOA*-NCOL*
Financial Assets	fna*	We use IVST+ IVAO. If either is missing, they are set to zero

Name	Abbreviation	Construction
Financial Liabilities	fnl*	We use DEBT*+PSTK*. If PSTK* is missing, it is set to zero
Net Financial Assets	nfna*	We use FNA*-FNL*
Operating Assets	oa*	We use COA*+NCOA*
Operating Liabilities	ol*	We use COL*+NCOL*
Net Operating Assets	noa*	We use OA*-OL*
Long-Term NOA	lnoa*	PPENT + INTAN + AO - LO + DP
Property Plant and Equipment Less Inventories	ppeinv*	PPEGT + INVT
Ortiz-Molina and Phillips Liquidity	aliq*	CHE + 0.75• COA* + 0.5(AT* - CA* - INTAN). If INTAN is missing, we set it to zero
Market Based		
Market Equity	me	We use the market equity for the stock we deem to the primary security of the firm. Importantly, we do not align the market value with the end of the fiscal period. Instead, we update the market value on a monthly basis and align it with the most recently available accounting characteristic
Market Enterprise Value	mev*	We use ME_COMPANY + NETDEBT* • FX*
Market Assets	mat*	We use AT* • FX + BE* • FX + ME_COMPANY
Accruals		
Operating Accruals	oacc*	We prefer NI*-OANCF. If that is unavailable, we use the yearly change in COWC*+the yearly change in NNCOA*
Total Accruals	tacc*	We use OACC* + the yearly change in NFNA*
Operating Cash Flow	ocf*	We prefer to use OANCF. If this is unavailable, we use NI*-OACC*. If this is unavailable, we use NI* + DP - WCAPT. If WCAPT is missing, we use 0.
Quarterly Operating Cash Flow	ocf_qtr*	We use OANCFQ. If this is unavailable, then we use IBQ + DPQ - WCAPTQ. If WCAPTQ is unavailable, we set it to
Cash Based Operating Profitability	cop*	We prefer EBITDA*+XRD-OACC*. If XRD is unavailable, we set it to zero
Other		
Employees in Thousands	emp	Compustat item EMP

Table E.9: Accounting Characteristics

Name	Abbreviation	Construction
Growth - Percentage[46]		
Asset Growth 1yr	at_gr1	$\frac{AT^*_t}{AT^*_{t-12}} - 1$
Sales Growth 1yr	sale_gr1	$\frac{SALE^*_t}{SALE^*_{t-12}} - 1$
Sales Growth 3yr	sale_gr3	$\frac{SALE^*_t}{SALE^*_{t-36}} - 1$
Total Debt Growth 3yr	debt_gr3	$\frac{DEBT^*_t}{DEBT^*_{t-36}} - 1$

[46]This refers to all variables with a suffix of "_gr1" or "_gr3". The variables are percentage growth in the accounting variables before the suffix. The number in the suffix refers to either 1 or 3 year growth. For all variables, we only take the percentage growth if the denominator is above zero.

Name	Abbreviation	Construction
Growth - Changed Scaled by Total Assets		
Inventory Change 1yr	inv_gr1a	$\frac{INV_t - INV_{t-12}}{AT^*_t}$
Investment and Advances Change 1yr	lti_gr1a	$\frac{LTI_t - LTI_{t-12}}{AT^*_t}$
Current Operating Assets Change 1yr	coa_gr1a	$\frac{COA^*_t - COA^*_{t-12}}{AT^*_t}$
Current Operating Liabilities Change 1yr	col_gr1a	$\frac{COL^*_t - COL^*_{t-12}}{AT^*_t}$
Non-Current Operating Assets Change 1yr	ncoa_gr1a	$\frac{NCOA^*_t - NCOA^*_{t-12}}{AT^*_t}$
Non-Current Operating Liabilities Change 1yr	ncol_gr1a	$\frac{NCOL^*_t - NCOL^*_{t-12}}{AT^*_t}$
Net Non-Current Operating Assets Change 1yr	nncoa_gr1a	$\frac{NNCOA^*_t - NNCOA^*_{t-12}}{AT^*_t}$
Net Operating Assets Change 1yr	noa_gr1a	$\frac{NOA^*_t - NOA^*_{t-12}}{AT^*_t}$
Financial Liabilities Change 1yr	fnl_gr1a	$\frac{FNL^*_t - FNL^*_{t-12}}{AT^*_t}$
Net Financial Assets Change 1yr	nfna_gr1a	$\frac{NFNA^*_t - NFNA^*_{t-12}}{AT^*_t}$
Effective Tax Rate Change 1yr	tax_gr1a	$\frac{TAX_t - TAX_{t-12}}{AT^*_t}$
Profit Margins		

Name	Abbreviation	Construction
Operating Profit Margin after Depreciation	ebit_sale	$\frac{EBIT^*_t}{SALE^*_t}$
Return on Assets		
Gross Profit scaled by Assets	gp_at	$\frac{GP^*_t}{AT^*_t}$
Cash Based Operating Profitability scaled by Assets	cop_at	$\frac{COP^*_t}{AT^*_t}$
Return on Book Equity		
Operating Profit to Equity scaled by BE	ope_be	$\frac{OPE^*_t}{BE^*_t}$
Net Income scaled by BE	ni_be	$\frac{NI^*_t}{BE^*_t}$
Return on Invested Capital		
Operating Profit after Depreciation scaled by BEV	ebit_bev	$\frac{EBIT^*_t}{BEV^*_t}$
Issuance		
Net Issuance scaled by Assets	netis_at	$\frac{NETIS^*_t}{AT^*_t}$
Equity Net Issuance scaled by Assets	eqnetis_at	$\frac{EQNETIS^*_t}{AT^*_t}$
Net Debt Issuance scaled by Assets	dbnetis_at	$\frac{DBNETIS^*_t}{AT^*_t}$
Accruals		
Operating Accruals	oaccruals_at	$\frac{OACC^*_t}{AT^*_t}$

Name	Abbreviation	Construction
Percent Operating Accruals	oaccruals_ni	$\frac{OACC^*_t}{\lvert NIX^*_t\rvert}$
Total Accruals	taccruals_at	$\frac{TACC^*_t}{AT^*_t}$
Percent Total Accruals	taccruals_ni	$\frac{TACC^*_t}{\lvert NIX^*_t\rvert}$
Net Operating Asset to Total Assets	noa_at	$\frac{NOA^*_t}{AT^*_t}$
Financial Soundness Ratios		
Operating Leverage	opex_at	$\frac{OPEX^*_t}{AT^*_t}$
Activity/Efficiency Ratios		
Asset Turnover	at_turnover	$\frac{SALE^*_t}{(AT^*_t + AT^*_{t-12})/2}$
Miscellaneous		
Sales scaled by BEV	sale_bev	$\frac{SALE^*_t}{BEV^*_t}$
R&D scaled by Sales	rd_sale	$\frac{XRD_t}{SALE^*_t}$
Balance Sheet Fundamental to Market Equity		
Book Equity scaled by Market Equity	be_me	$\frac{BE^*_t}{ME_t}$
Total Assets scaled by Market Equity	at_me	$\frac{AT^*_t}{ME_t}$
Income Fundamentals to Market Equity		

Name	Abbreviation	Construction
Net Income scaled by ME	ni_me	$\frac{NI^*_t}{ME_t}$
Sales scaled by ME	sale_me	$\frac{SALE^*_t}{ME_t}$
Operating Cash Flow scaled by ME	ocf_me	$\frac{OCF^*_t}{ME_t}$
Free Cash Flow scaled by ME	fcf_me	$\frac{FCF^*_t}{ME_t}$
R&D scaled by ME	rd_me	$\frac{XRD_t}{ME_t}$
Balance Sheet Fundamentals to Market Enterprise Value		
Book Enterprise Value scaled by MEV	bev_mev	$\frac{BEV^*_t}{MEV^*_t}$
Equity Payout/Issuance to Market Equity		
Net Equity Payout scaled by ME	eqpo_me	$\frac{EQPO^*_t}{ME_t}$
Equity Net Payout scaled by ME	eqnpo_me	$\frac{EQNPO^*_t}{ME_t}$
Income Fundamentals to Market Enterprise Value		
Operating Profit before Depreciation scaled by MEV	ebitda_mev	$\frac{EBITDA^*_t}{MEV^*_t}$
New Variables not in HXZ		
Operating Cash Flow scaled by Assets	ocf_at	$\frac{OCF^*_t}{AT^*_t}$

Name	Abbreviation	Construction
Operating Cash Flow to Assets 1 yr Change	ocf_at_chg1	$OCF\_AT_t - OCF\_AT_{t-12}$
New Variables from HXZ		
Cash and Short Term Investments scaled by Assets	cash_at	$\frac{CASH_t}{AT^*_t}$
Number of Consecutive Earnings Increases	ni_inc8q	Count number of earnings increases over past 8 quarters
Change in Property, Plant and Equipment Less Inventories scaled by lagged Assets	ppeinv_gr1a	$\frac{PPEINV^*_t - PPEINV^*_{t-12}}{AT^*_{t-12}}$
Change in Long-Term NOA scaled by average Assets	lnoa_gr1a	$\frac{LNOA^*_t - LNOA^*_{t-12}}{AT^*_t - AT^*_{t-12}}$
CAPX 1 year growth	capx_gr1	$\frac{CAPX_t}{CAPX_{t-12}} - 1$
CAPX 2 year growth	capx_gr2	$\frac{CAPX_t}{CAPX_{t-24}} - 1$
CAPX 3 year growth	capx_gr3	$\frac{CAPX_t}{CAPX_{t-36}} - 1$
Change in Short-Term Investments scaled by Assets	sti_gr1a	$\frac{IVST_t - IVST_{t-12}}{AT^*_t}$
Quarterly Income scaled by BE	niq_be	$\frac{NI\_QTR^*_t}{BE^*_{t-3}}$
Change in Quarterly Income scaled by BE	niq_be_chg1	$NIQ\_BE_t - NIQ\_BE_{t-12}$
Quarterly Income scaled by AT	niq_at	$\frac{NI\_QTR^*_t}{AT^*_{t-3}}$

Name	Abbreviation	Construction
Change in Quarterly Income scaled by AT	niq_at_chg1	$NIQ\_AT_t - NIQ\_AT_{t-12}$
Quarterly Sales Growth	saleq_gr1	$\frac{SALE\_QTR^*_t}{SALE\_QTR^*_{t-12}} - 1$
R&D Capital-to-Assets	rd5_at	$\frac{\sum_{n=0}^{4}(1 - .2 \bullet n)(XRD_{t-12*n})}{AT^*_t}$
Age	age	Age of the firms in months
Change Sales minus Change Inventory	dsale_dinv	$CHG\_TO\_EXP(SALE^*_t) - CHG\_TO\_EXP(INV_t)$
Change Sales minus Change Receivables	dsale_drec	$CHG\_TO\_EXP(SALE^*_t) - CHG\_TO\_EXP(REC_t)$
Change Gross Profit minus Change Sales	dgp_dsale	$CHG\_TO\_EXP(GP^*_t) - CHG\_TO\_EXP(SALE^*_t)$
Change Sales minus Change SG&A	dsale_dsga	$CHG\_TO\_EXP(SALE^*_t) - CHG\_TO\_EXP(XSGA_t)$
Earnings Surprise	saleq_su	$SUR(SALE\_QTR^*)$
Revenue Surprrise	niq_su	$SUR(NI\_QTR^*)$
Total Debt scaled by ME	debt_me	$\frac{DEBT^*_t}{ME_t}$
Net Debt scaled by ME	netdebt_me	$\frac{NETDEBT^*_t}{ME_t}$
Abnormal Corporate Investment	capex_abn	$\frac{CAPX\_SALE^*_t}{(CAPX\_SALE^*_{t-12} + CAPX\_SALE^*_{t-24} + CAPX\_SALE^*_{t-36})/3} - 1$

Name	Abbreviation	Construction
Inventory Change 1 yr	inv_gr1	$\frac{INV_t}{INV_{t-12}} - 1$
Book Equity Change 1 yr scaled by Assets	be_gr1a	$\frac{BE^*_t - BE^*_{t-12}}{AT^*_t}$
Ball Operating Profit to Assets	op_at	$\frac{OP^*_t}{AT^*_t}$
Earnings before Tax and Extraordinary Items to Net Income Including Extraordinary Items	pi_nix	$\frac{PI^*_t}{NIX^*_t}$
Ball Operating Profit scaled by lagged Assets	op_atl1	$\frac{OP^*_t}{AT^*_{t-12}}$
Operating Profit scaled by lagged Book Equity	ope_bel1	$\frac{OPE^*_t}{BE^*_{t-12}}$
Gross Profit scaled by lagged Assets	gp_atl1	$\frac{GP^*_t}{AT^*_{t-12}}$
Cash Based Operating Profitability scaled by lagged Assets	cop_atl1	$\frac{COP^*_t}{AT^*_{t-12}}$
Book Leverage	at_be	$\frac{AT^*_t}{BE^*_t}$
Operating Cash Flow to Sales Quarterly Volatility	ocfq_saleq_std	$SDEV\_16Q\left(\frac{OCF\_QTR^*_t}{SALE\_QTR^*_t})\right)$
Liquidity scaled by lagged Assets	aliq_at	$\frac{ALIQ^*_t}{AT^*_{t-12}}$

Name	Abbreviation	Construction
Liquidity scaled by lagged Market Assets	aliq_mat	$\frac{ALIQ^*_t}{MAT^*_{t-12}}$
Tangibility	tangibility	$\frac{CASH_t + 0.715 \bullet REC_t + 0.547 \bullet INV_t + 0.535 \bullet PPEG_t}{AT^*_t}$
Equity Duration	eq_dur	Following Dechow, Sloan and Soliman (2004)
Piotroski F-Score	f_score	Following Piotroski (2000)
Ohlson O-Score	o_score	Following Ohlson (1980)
Altman Z-Score	z_score	Following Altman (1968)
Kaplan-Zingales Index	kz_index	Following Kaplan and Zingales (1997)
Intrinsic ROE	intrinsic_value	Following Frankel and Lee (1998)
Sales scaled by Employees Growth 1 yr	sale_emp_gr1	$\frac{SALE\_EMP_t}{SALE\_EMP_{t-12}} - 1$
Employee Growth 1 yr	emp_gr1	$\frac{EMP_t - EMP_{t-12}}{0.5 \bullet EMP_t + 0.5 \bullet EMP_{t-12}}$
Earnings Variability	earnings_variability	$\frac{SDEV\_60M\left(\frac{NI^*_t}{AT^*_{t-12}}\right)}{SDEV\_60M\left(\frac{OCF^*_t}{AT^*_{t-12}}\right)}$
1 yr lagged Net Income to Assets	ni_ar1	$\frac{NI^*_{t-12}}{AT^*_{t-12}}$
Net Income Idiosyncratic Volatility	ni_ivol	Following Francis et al. (2004)

E.3 Market Based Characteristics

E.3.1 Datasets

- CRSP.MSF
- CRSP.DSF
- COMP.SECD
- COMP.G_SECD
- COMP.FUNDQ
- COMP.FUNDA
- COMP.SECM
- COMP.SECURITY
- COMP.G_SECURITY

E.3.2 Market Variables

The abbreviation is used to refer to the accounting variable. A suffix of '*' indicates that we have altered the original Compustat item to increase the coverage. The characteristic name will reflect the accounting name except the '*' suffix. As an example, 'gp_at' is gross profit scaled by assets. In general, we will refer to Compustat variables using capital letters. We use the CRSP Market Variable values if they are available, and if they are not, we use the Compustat Market Variables.

Table E.10: Market Variables

Name	Abbreviation	Construction
CRSP Variables[47]		
Share Adjustment Factor	adjfct*	We use CFACSHR
Shares	shares*	We use SHROUT/100
Price	prc*	We use |PRC|
Adjusted Proce	prc_adj*	We use PRC*•ADJFCT*
Market Equity	me*	We use PRC*•SHARES*
Dollar Volume	dolvol*	We use VOL•PRC*
Return	RET*	We use RET
Excess Return	ret_exc*	We use (RET*-T30RET)/21. If T30RET is unavailable, we use RF. If the return is a daily return rather than a monthly return, the RET - T30RET is divided by 1 rather than 21.
Cumulative Return	ri*	This is the cumulative return estimated from RET*
Monthly Dividend	div_tot*	We use (RET -RETX)•lag(PRC*)•(CFACSHR/lag(CFACSHR))
Compustat Variables		
Share Adjustment Factor	adjfct*	We use AJEXDI
Shares	shares*	We use CSHOC/1000000
Price	prc*	We use PRC_LOCAL*•FX [48]
Local Price	prc_local*	We use PRCCD
Market Equity	me*	We use PRC*•SHARES*
Dollar Volume	dolvol*	We use CSHTRD•PRC*
Return	RET*	We use RET_LOCAL*•FX
Cumulative Return - Local	ri_local*	We use PRC_LOCAL*• TRFD/AJEXDI
Local Return	ret_local*	We use RI_LOCAL*/lag(RI_LOCAL*) - 1
Cumulative Return	ri*	RI_LOCAL* • FX*
Monthly Dividend	div_tot*	We use DIV • FX*. If DIV is missing, we set it to zero
Asset Pricing Factors		
Excess Market Return	mktrf*	Country specific market return
High Minus Low	hml*	Country specific factor following Fama and French (1993) and using breakpoints from non-micro cap stocks within the country
Small Minus Big ala Fama-French	smb_ff*	Average of small portfolios minus average of large portfolios from hml*
Return on Equity	roe*	Country specific factor following Hou, Xue and Zhang (2015) and using breakpoints from non-micro cap stocks within the country. We use double sorts on return on equity and size rather than triple sorts with investment, due to the limited number of stocks in some international markets.
Investment	inv*	Country specific factor following Hou, Xue and Zhang (2015) and using breakpoints from non-micro cap stocks within the country. We use double sorts on investment and size rather than triple sorts with return on equity, due to the limited number of stocks in some international markets

[47] lag is a lag function where lag(x) is the value of x from the previous time period

[48] FX scales the price to USD

Name	Abbreviation	Construction
Small Minus Big ala Hou et al	smb_hxz*	Average of small portfolios minus average of large portfolios from roe* and inv*
Market Volatility for Each Stock	_mktvol_zd*	$SDEV\_zD(MKTRF^*_t)$ [49]

Table E.11: Market Characteristics

Name	Abbreviation	Construction
Size Based Measures		
Market Equity	market_equity	ME^*_t
Total Dividend Paid to Market Equity		
Dividend to Price - 12 Months	div12m_me	$\frac{\sum_{n=0}^{11} DIV\_TOT^*_{t-n} \bullet SHARES^*_{t-n}}{ME^*_t}$
Change in Shares Outstanding		
Change in Shares - 12 Month	chcsho_12m	$\frac{SHARES^*_t \bullet ADJFCT^*_t}{SHARES^*_{t-12} \bullet ADJFCT^*_{t-12}} - 1$
Net Equity Payout		
Net Equity Payout - 12 Month	eqnpo_12m	$log\left(\frac{RI^*_t}{RI^*_{t-12}}\right) - log\left(\frac{ME^*_t}{ME^*_{t-12}}\right)$
Momentum/Reversal		
Short Term Reversal	ret_1_0	$\frac{RI^*_t}{RI^*_{t-1}} - 1$
Momentum 1-3 Months	ret_3_1	$\frac{RI^*_{t-1}}{RI^*_{t-3}} - 1$
Momentum 1-6 Months	ret_6_1	$\frac{RI^*_{t-1}}{RI^*_{t-6}} - 1$
Momentum 1-9 Months	ret_9_1	$\frac{RI^*_{t-1}}{RI^*_{t-9}} - 1$

[49] Must have enough non-missing values of stock to be estimated

Name	Abbreviation	Construction
Momentum 1-12 Months	ret_12_1	$\frac{RI^*_{t-1}}{RI^*_{t-12}} - 1$
Momentum 7-12 Months	ret_12_7	$\frac{RI^*_{t-7}}{RI^*_{t-12}} - 1$
Momentum 12-60 Months	ret_60_12	$\frac{RI^*_{t-12}}{RI^*_{t-60}} - 1$
Seasonality		
1 Year Annual Seasonality	seas_1_1an	Return in month t-12
2 - 5 Year Annual Seasonality	seas_2_5an	Average return over annual lags from year t-2 to t-5
6 - 10 Year Annual Seasonality	seas_6_10an	Average return over annual lags from year t-6 to t-10
11 - 15 Year Annual Seasonality	seas_11_15an	Average return over annual lags from year t-11 to t-15
16 - 20 Year Annual Seasonality	seas_16_20an	Average return over annual lags from year t-16 to t-20)
1 Year Non-Annual Seasonality	seas_1_1na	Average return from month t-1 to t-11
2 - 5 Year Non-Annual Seasonality	seas_2_5na	Average return over non-annual lags from year t-2 to t-5
6 - 10 Year Non-Annual Seasonality	seas_6_10na	Average return over non-annual lags from year t-6 to t-10
11 - 15 Year Non-Annual Seasonality	seas_11_15na	Average return over non-annual lags from year t-11 to t-15
16 - 20 Year Non-Annual Seasonality	seas_16_20na	Average return over non-annual lags from year t-16 to t-20
Combined Accounting and Market Based Characteristics		
Let e_t be the residuals of a cross-sectional regression of RET_EXC_t on $MKTRF_t$, SMB_FF_t and HML_t		
60 Month CAPM Beta	beta_60m	$\frac{COVAR\_60M(RET^*_t, MKTRF^*_t)}{VARC\_60M(MKTRF^*_t)}$
Performance Based Mispricing	mispricing_perf[50]	$\frac{1}{4}(O\_SCORE^{r01}_t + RET\_12\_1^{r01}_t + GP\_AT^{r01}_t + NIQ\_AT^{r01}_t)$
Management Based Mispricing	mispricing_mgmt	$\frac{1}{6}(CHCSHO\_12M^{r01}_t + EQNPO\_12M^{r01}_t + OACCRUALS\_AT^{r01}_t + NOA\_AT^{r01}_t + AT\_GR1^{r01}_t + PPEINV\_GR1A^{r01}_t)$
Residual Momentum - 6 Month	resff3_6_1	$-1 + \prod_{n=1}^{6} 1 + e_{t-n}$
Residual Momentum - 12 Month	resff3_12_1	$-1 + \prod_{n=1}^{12} 1 + e_{t-n}$
Daily Market Data[51]		
Let ϵ_t be the residuals of a cross-sectional regression of RET_EXC_t on $MKTRF_t$		

[50] A rank characteristic has the value of that characteristics rank with respect to other companies' same characteristic of the same month and country scaled [0, 1]. This is identified with a "r01" superscript.

[51] Many of the variables in this section are estimated using rolling windows of data, and the variables are estimated using a variety of window lengths: 21, 126, 252 and 1260 days. In this section, I refer to the

Name	Abbreviation	Construction
Let σ_t be the residuals of a cross-sectional regression of RET_EXC_t on $MKTRF_t$, SMB_HXZ_t, ROE_t, and INV_t		
Return Volatility	rvol_zd	$SDEV\_zD(RET\_EXC^*{}_t)$
Maximum Return	rmax1_zd	$MAX1\_zD(RET^*{}_t)$
Mean Maximum Return	rmax5_zd	$\frac{1}{5}\sum_{n=1}^{5} X_n, \;\; X_n \in MAX5\_zD(RET^*)$
Return Skewness	rskew_zd	$SKEW\_zD(RET\_EXC^*{}_t)$
Price-to-High	prc_highprc_zd	$\frac{PRC\_ADJ^*{}_t}{MAX1\_zD(PRC\_ADJ^*{}_t)}$
Amihud (2002) Measure	ami_zd	$MEAN\_zD\left(\frac{\lvert RET^*{}_t\rvert}{DOLVOL^*{}_t}\right) * 1000000$
CAPM Idiosyncratic Vol.	ivol_capm_zd	$SDEV\_zD(\epsilon_t)$
CAPM Idiosyncratic Skewness	iskew_capm_zd	$SKEW\_zD(\epsilon_t)$
Coskewness	coskew_zd[52]	$\frac{MEAN\_zD(\epsilon_t \bullet MKTRF\_DM_t^2)}{\sqrt{MEAN\_zD(\epsilon_t^2) \bullet MEAN\_zD(MKTRF\_DM_t^2)}}$
Fama and French Idiosyncratic Vol.	ivol_ff3_zd	$SDEV\_zD(e_t)$

number of days as m as a proxy for any of the possible window lengths.

[52] $MKTRF\_DM_t = MKTRF^*{}_t - MEAN\_zD(MKTRF^*{}_t)$

Name	Abbreviation	Construction
Fama and French Idiosyncratic Skewness	iskew_ff3_zd	$SKEW\_zD(e_t)$
Hou, Xue and Zhang Idiosyncratic Vol.	ivol_hxz4_zd	$SDEV\_zD(\sigma_t)$
Hou, Xue and Zhang Idiosyncratic Skewness	iskew_hxz4_zd	$SKEW\_zD(\sigma_t)$
Dimson Beta	beta_dimson_zd	Following Dimson (1979)
Downside Beta	betadown_zd	Coefficient from regression of $RET\_EXC^*{}_t$ on $MKTRF^*{}_t$ when $MKTRF^*{}_t < 0$
Zero Trades	zero_trades_zd	Number of days with zero trades over period. In case of equal number of zero trading days, turnover_zd will decide on the rank following Liu (2006)
Turnover	turnover_zd	$MEAN\_zD(\frac{TVOL^*{}_t}{SHARES^*{}_t * 1000000})$
Turnover Volatility	turnover_var_zd	$\frac{SDEV\_zD\left(\frac{TVOL^*{}_t}{SHARES^*{}_t * 1000000}\right)}{TURNOVER\_zD_t}$
Dollar Volume	dolvol_zd	$MEAN\_zD(DOLVOL^*{}_t)$
Dollar Volume Volatility	turnover_var_zd	$\frac{SDEV\_zD(DOLVOL^*{}_t)}{DOLSDEV\_zD_t}$
Correlation to Market	corr_zd	The correlation between $RET\_EXC^*{}_3l = RET\_EXC^*{}_t + RET\_EXC^*{}_{t-1} + RET\_EXC^*{}_{t-2}$ and $MKT\_EXC\_3l = MKTRF^*{}_t + MKTRF^*{}_{t-1} + MKTRF^*{}_{t-2}$
Betting Against Beta	betabab_1260d	$\frac{CORR\_1260d_t \bullet RVOL\_252d_t}{\_MKTVOL\_252d^*{}_t}$
Max Return to Volatility	rmax5_rvol_21d	$\frac{RMAX5\_21d_t}{RVOL\_252d_t}$
21 Day Bid-Ask High-Low	bidaskhl_21d	High-low bid ask estimator created using code from Corwin and Schultz (2012)
Quality Minus Junk		

Name	Abbreviation	Construction
Quality Minus Junk - Profit	qmj_prof	$ZV\Big(ZV(GP\_AT_t) + ZV(NI\_BE_t) + ZV(NI\_AT_t) + ZV(OCF\_AT_t) + ZV(GP\_SALE^*{}_t) + ZV(OACCRUALS\_AT_t)\Big)$
Quality Minus Junk - Growth	qmj_growth	$ZV\Big(ZV(GPOA\_CH5_t) + ZV(ROE\_CH5_t) + ZV(ROA\_CH5_t) + ZV(CFOA\_CH5_t) + ZV(GMAR\_CH5_t)\Big)$
Quality Minus Junk - Safety	qmj_safety	$ZV\Big(ZV(BETABAB\_1260d_t) + ZV(DEBT\_AT_t) + ZV(O\_SCORE_t) + ZV(Z\_SCORE_t) + ZV(\_EVOL_t)\Big)$
Quality Minus Junk	qmj	$\frac{QMJ\_PROF_t + QMJ\_GROWTH_t + QMJ\_SAFETY_t}{3}$