Internet Appendix
Factors and Risk Premia in Individual International Stock Returns

January 15, 2020

In this Internet Appendix, we provide all estimation and inference methodologies used in this paper. We provide additional details and results that complement those in the main text.

Appendix A details the estimation methodology and inference tools. Appendix B details our data construction. Appendix C presents results on beta determinants for the mixed regional four- and q-factor models.

A Estimation and test methodology

We detail in this section the estimation and test methodology with common instruments $Z_{c,t}$ for stocks of country c and stock-specific instruments $Z_{c,t}(\gamma_{i,c})$ where $\gamma_{i,c}$ corresponds to a draw of asset i in country c, $i = 1, ..., n_c$.

A major impediment to testing the IAPT with multiple factors, time-varying factor exposures and risk premia is the proliferation of parameters to estimate. To address this problem, we extend the methodology of Gagliardini, Ossola, and Scaillet (2016) in the following way. For each stock, we use an automatic procedure to select the most important instruments while making sure that model specification is consistent with no-arbitrage conditions for *each* stock. Then, we adjust the asset pricing test to account for the different number of parameters across stocks. We detail each step below.

A.1 Time-series regressions

Our specification choices for factor exposures and factor risk premia (Equations (5), (6), and (7) in the main text) combined with the asset pricing restrictions, $a_{c,t}(\gamma) = b_{c,t}(\gamma)\nu_{c,t}$, imply that a stock intercept is,

$$a_{i,c,t} = Z'_{c,t-1}B'_{i,c}\left(\Lambda_c - F_c\right)Z_{c,t-1} + Z'_{i,c,t-1}C'_{i,c}\left(\Lambda_c - F_c\right)Z_{c,t-1},$$

using the simplifying notations $a_{c,t}(\gamma_{i,c}) = a_{i,c,t}$, $B_c(\gamma_{i,c}) = B_{i,c}$, $Z_{c,t}(\gamma_{i,c}) = Z_{i,c,t}$, and $C_c(\gamma_{i,c}) = C_{i,c}$.

To avoid the curse of dimensionality, we impose some structure and set some elements in $B_{i,c}$ and $C_{i,c}$ to zero. Let $\mathbb{I}_{B_{i,c}}$ and $\mathbb{I}_{C_{i,c}}$ be a K-by-p and a K-by-q indicator matrices whose elements are equal to one if the corresponding elements in $B_{i,c}$ and $C_{i,c}$ are not zero. The i subscript indicates that the indicator matrices can change across stocks and we explain in Section A.2 how they are selected.

We further define $\tilde{\mathbb{I}}_{B_{i,c}}$ as the p-vector whose j^{th} element is equal to one if at least one element in the j^{th} column of $B_{i,c}$ is not zero and equal to zero otherwise. Let $\tilde{p}_i = \tilde{\mathbb{I}}'_{B_{i,c}}\iota_p$ be the number of columns in $B_{i,c}$ with at least one non-zero element where ι_p be a p-vector of ones.

We extend the GOS methodology in our high-dimensional international setting using these indicator matrices to reduce the number of parameters to estimate. One possibility would be to compute the Hadamard products, $\mathbb{I}_{B_{i,c}} \circ B_{i,c}$ and $\mathbb{I}_{C_{i,c}} \circ C_{i,c}$ and use them in the time-series regressions. However, this approach based on simple element-by-element products would create regressors whose values are all zeros and would prevent running the time-series regressions. Instead, to get feasible least-squares estimations we use these indicator matrices to select the required regressors as follows.

We impose some structure on the use of common and stock-specific instruments. First, all elements in the first column of $\mathbb{I}_{B_{i,c}}$ are equal to one, i.e., we always use a constant to model factor exposures and $\tilde{p}_i \geq 1$. Those exposures correspond to the time-invariant factor contributions. Second, at least one element in each column of $\mathbb{I}_{C_{i,c}}$ is equal to one to ensure all stock-specific instruments are relevant for at least one factor. We do not impose *a priori* any structure on the matrix $\Lambda_c - F_c$. Our approach is general as it only requires specifying the elements in $B_{i,c}$ and $C_{i,c}$ as long as the above two restrictions are respected.

To simplify the notation, we define the $d_{1,i} = p(p-1)/2 + \tilde{p}_i + pq$ vector of predetermined variables

$$x_{i,c,t,1} = \left(\text{vech}(X_t)'\tilde{D}_i, Z'_{c,t-1} \otimes Z'_{i,c,t-1}\right)', \qquad (1)$$

where the matrix X_t has typical diagonal elements $X_{k,k,t} = Z^2_{c,t-1,k}$ and off-diagonal elements $X_{k,l,t} = 2Z_{c,t-1,k}Z_{c,t-1,l}$, the matrix $\tilde{D}_i$ is obtained by removing columns in

$$diag\left(\text{vech}\left[diag\left(\tilde{\mathbb{I}}_{B_{i,c}}\right) + \iota_p \iota_p' - I_p\right]\right)$$

for which all elements are equal to 0, and where I_p is the identity matrix of size p. *** vec vech def

In addition, we define the $d_{2,i} = (n_{\mathbb{I}_{B_{i,c}}} + n_{\mathbb{I}_{C_{i,c}}})$-vector of factors scaled by $Z_{c,t-1}$ (scaled factors) and by $Z_{i,c,t-1}$

$$x_{i,c,t,2} = \left[\left(f_t' \otimes Z'_{c,t-1}\right)\tilde{B}_i', \left(f_t' \otimes Z'_{i,c,t-1}\right)\tilde{C}_i'\right]', \qquad (2)$$

where $n_{\mathbb{I}_{B_{i,c}}}$ and $n_{\mathbb{I}_{C_{i,c}}}$ are respectively the number of non-zero elements in $B_{i,c}$ and $C_{i,c}$. In this equation, the $n_{\mathbb{I}_{B_{i,c}}}$-by-Kp matrix $\tilde{B}_i$ is obtained by removing the rows in $diag\left(vec\left[\mathbb{I}'_{B_{i,c}}\right]\right)$ for which all elements are equal to zero. Similarly, the $n_{\mathbb{I}_{C_{i,c}}}$-by-Kq matrix $\tilde{C}_i$ is obtained by removing the rows in $diag\left(vec\left[\mathbb{I}'_{C_{i,c}}\right]\right)$ for which all elements are equal to zero. The new definitions of $x_{i,c,t,1}$ and $x_{i,c,t,2}$ relying on the new matrices $\tilde{B}_i$ and $\tilde{C}_i$ clarify the differences with the GOS setting.

Then, we can use the compact notation with the d_i-vector $x_{i,c,t} = \left(x'_{i,c,t,1}, x'_{i,c,t,2}\right)'$ with $d_i = d_{1,i} + d_{2,i}$,

$$r_{i,c,t} = \beta'_{i,c} x_{i,c,t} + \varepsilon_{i,c,t}, \qquad (3)$$

with $\beta_{i,c} = \left(\beta'_{i,c,1}, \beta'_{i,c,2}\right)'$, and

$$\begin{aligned}
\beta_{i,c,1} &= \left(\left(X' N_p \left[(\Lambda_c - F_c)' \otimes I_p\right] \tilde{B}_i' \tilde{B}_i vec\left[B'_{i,c}\right]\right)', \left(\left[(\Lambda_c - F_c)' \otimes I_q\right] vec\left[C'_{i,c}\right]\right)'\right)', \\
\beta_{i,c,2} &= \left(\left(\tilde{B}_i vec\left[B'_{i,c}\right]\right)', \left(\tilde{C}_i vec\left[C'_{i,c}\right]\right)'\right)',
\end{aligned}$$

where $N_p = \frac{1}{2} D_p^+ \left(W_{p,p} + I_{p^2}\right)$, $W_{p,q}$ is the commutation matrix such that $\text{vec}(A') = W_{p,q}\text{vec}(A)$

for a p-by-q matrix A, and D_p^+ is the $p(p+1)/2$-by-p^2 matrix such that $\text{vech}(A) = D_p^+ \text{vec}(A)$. Here the dimension d_i of the vector $x_{i,c,t}$ depends on asset i while it does not in GOS. The dimension of their parameter vector $\beta_{i,c}$ is held fixed across stocks $d_i = d$.

To ensure that cross-sectional limits exist and are invariant to reordering of the assets, we introduce a sampling scheme as in GOS. Observable assets are random draws $i = 1, ..., n_c$ from an underlying population (Andrews, 2005). By random sampling, we get a standard random coefficient model (see, for example, Hsiao, 2003, Chapter 6).

We account for the unbalanced nature of the panel through a collection of indicator variables: we define $I_{i,c,t} = 1$ if the return of asset i in country c is observable at date t, and 0 otherwise (Connor and Korajczyk, 1987). The first pass consists in computing time-series OLS estimators

$$\hat{\beta}_{i,c} = \hat{Q}_{x,i,c}^{-1} \frac{1}{T_{i,c}} \sum_t I_{i,c,t} x_{i,c,t} r_{i,c,t},$$

for all stocks $i = 1, ..., n_c$, where $\hat{Q}_{x,i,c} = \dfrac{1}{T_{i,c}} \sum\limits_t I_{i,c,t} x_{i,c,t} x'_{i,c,t}$ and $T_{i,c} = \sum\limits_t I_{i,c,t}$.

A.2 Stock and instrument selection

The random sample size $T_{i,c}$ for stock i in country c can be small, and the inversion of matrix $\hat{Q}_{x,i,c}$ can be numerically unstable, possibly yielding unreliable estimates of $\beta_{i,c}$. Also, given that we use the cross-sectional ranks of characteristics as stock-specific instruments, we encounter many cases of multicollinearity. Consider for example a stock that remains among the largest stocks in its market during the sample period. Then its size cross-sectional rank, $Z_{i,c,t-1}^{size}$, is relatively constant and the interaction terms $Z_{i,c,t-1}^{size} f_{c,t}$ are highly correlated with $f_{c,t}$ for all factors in the regression.

To address this problem, we first keep only stocks with at least five years of monthly returns, $T_{i,c} \geq 60$. Then, we select instruments for the factor exposures to obtain a time-series regression that is not too badly conditioned. We follow these steps:

1. As a criterion we use the condition number which is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue of $\hat{Q}_{x,i,c}$, $CN\left(\hat{Q}_{x,i,c}\right) = \sqrt{eig_{\max}\left(\hat{Q}_{x,i,c}\right) / eig_{\min}\left(\hat{Q}_{x,i,c}\right)}$. A too large value of $CN\left(\hat{Q}_{x,i,c}\right)$ indicates multicollinearity problems and ill-conditioning (Belsley et al., 2004; Greene, 2008). In our empirical tests, we use a threshold of 50.

2. If the condition number for stock i is above 50, then we find the pair of regressors in $x_{i,c,t,2}$ that has the largest cross-correlation in absolute value. Of these two regressors, we remove the one with the lowest absolute correlation with $r_{i,c,t}$ and set its corresponding element in $\mathbb{I}_{B_{i,c}}$ or $\mathbb{I}_{C_{i,c}}$ to 0.

3. We check that the two conditions on the specification are respected, namely that the first column of $B_{i,c}$ are all selected and that there is at least one element selected in each of the columns of $C_{i,c}$. Otherwise, we keep both regressors and look for the next regressor pair with the highest absolute cross-correlation.

4. Given our new selection of regressors $x_{i,c,t,2}$, we compute the new matrix $\tilde{D}_i$ and use Equation (1) to obtain the new selected $x_{i,c,t,1}$. Therefore, our methodology selects instruments for the factors exposures and ensures that the time-series regression specification is consistent with no-arbitrage conditions.

We follow these steps until the condition number falls below 50. If there are no specification that respects the two specification conditions and has a low enough condition number, then we do not keep stock i. We then define the indicator variable $\mathbf{1}^{\chi}_{i,c}$ which takes a value of one if stock i is kept and zero otherwise.

In our empirical tests, we have found that this cleaning procedure produces better results in terms of time-series fit and pricing errors than selecting *a priori* the instruments to use for each factor.

A.3 Cross-sectional regressions

The second pass consists in computing a cross-sectional estimator of ν_c by regressing the $\hat{\beta}_{i,c,1}$s on the $\hat{\beta}_{i,c,3}$s keeping only the non-trimmed assets. First, we can write $\beta_{i,c,1}$ as

$$\beta_{i,c,1} = \beta_{i,c,3}\nu_c, \tag{4}$$

where

$$\begin{aligned} \beta_{i,c,3} &= \left(\left(\tilde{D}_i' N_p \left[B_{i,c}' \otimes I_p\right]\right)', \left[W_{p,q}\left(C_{i,c}' \otimes I_p\right)\right]'\right)', \\ \nu_c &= vec\left[\Lambda_c' - F_c'\right]. \end{aligned}$$

We obtain $\beta_{i,c,3}$ using the following identity,

$$\begin{aligned} \text{vec}(\beta_{i,c,3}') &= J_{a,i}\beta_{i,c,2}, \\ J_{a,i} &= \begin{pmatrix} J_{1,i} & 0 \\ 0 & J_{2,i} \end{pmatrix}, \\ J_{1,i} &= W_{p(p-1)/2+\tilde{p}_i,Kp}\left(I_{pK} \otimes \left(\tilde{D}_i' N_p\right)\right)\{I_K \otimes [(W_p \otimes I_p)\left(I_p \otimes vec\left[I_p\right]\right)]\}\tilde{B}_i', \\ J_{2,i} &= W_{pq,pK}\left(I_K \otimes [(I_p \otimes W_{p,q})(W_{p,q} \otimes I_p)(I_q \otimes \text{vec}(I_p))]\right)\tilde{C}_i'. \end{aligned}$$

We use a Weighted Least Squares (WLS) approach,

$$\hat{\nu}_c^{WLS} = \hat{Q}_{\beta_3}^{-1} \frac{1}{n_c} \sum_i \hat{\beta}'_{i,c,3} \hat{w}_{i,c} \hat{\beta}_{i,c,1}, \tag{5}$$

where $\hat{Q}_{\beta_3} = \frac{1}{n_c} \sum_i \hat{\beta}'_{i,c,3} \hat{w}_{i,c} \hat{\beta}_{i,c,3}$ and $\hat{w}_{i,c} = \mathbf{1}_{i,c}^{\chi} (diag[\hat{v}_{i,c}])^{-1}$ are the weights.

The terms $v_{i,c} = \tau_{i,c} C'_{\nu_c,i} Q_{x,i,c}^{-1} S_{ii,c} Q_{x,i,c}^{-1} C_{\nu_c,i}$ are the asymptotic variances of the standardized errors $\sqrt{T}\left(\hat{\beta}_{i,c,1} - \hat{\beta}_{i,c,3}\nu_c\right)$ in the cross-sectional regression for large T, where $\tau_{i,c} = E[I_{i,c,t}|\gamma_{i,c}]^{-1}$, $Q_{x,i,c} = E\left[x_{i,c,t}x'_{i,c,t}\right]$, $S_{ii,c} = E\left[\varepsilon_{i,c,t}^2 x_{i,c,t} x'_{i,c,t}|\gamma_{i,c}\right]$, $C_{\nu_c,i} = (E'_{1,i} - (I_{d_{1,i}} \otimes \nu'_c) J_{a,i} E'_{2,i})'$, $E_{1,i} = \left(I_{d_{1,i}}, \mathbf{0}_{d_{1,i},d_{2,i}}\right)'$, $E_{2,i} = \left(\mathbf{0}_{d_{2,i},d_{1,i}}, I_{d_{2,i}}\right)'$, and $\mathbf{0}_{d_{1,i},d_{2,i}}$ is a $d_{1,i}$-by-$d_{2,i}$ matrix of zeros.

To operationalize this WLS approach, we first estimate $\hat{\nu}_c^{OLS}$ by OLS using unit weights $\hat{w}_{i,c} = 1$. We then use the estimates $\tau_{i,c} = \frac{T_c}{T_{i,c}}$, $C_{\hat{\nu}_c,i} = (E'_{1,i} - (I_{d_{1,i}} \otimes \hat{\nu}_c^{OLS'}) J_{a,i} E'_{2,i})'$, $\hat{S}_{ii,c} = \frac{1}{T_{i,c}} \sum_t I_{i,c,t} \hat{\varepsilon}_{i,c,t}^2 x_{i,c,t} x'_{i,c,t}$, and $\hat{\varepsilon}_{i,c,t} = r_{i,c,t} - \hat{\beta}'_{i,c} x_{i,c,t}$ to estimate $\hat{\nu}_c^{WLS}$ by WLS.[1]

The distribution of $\hat{\nu}_c^{WLS}$ is

$$\sqrt{n_c T_c}\left(\hat{\nu}_c^{WLS} - \frac{1}{T_c}\hat{B}_{\nu_c} - \nu_c\right) \Rightarrow N\left(0, \Sigma_{\nu_c}\right), \tag{6}$$

where the presence of the bias term $\hat{B}_{\nu_c}$ comes from the well-known Error-In-Variable problem, that is, factor exposures are estimated with errors in the first step time-series regressions. We report the expressions for the bias term $\hat{B}_{\nu_c}$ and the estimation methodology for the covariance matrix Σ_{ν_c} in the following two sections.

To obtain estimates of the time-varying risk premia, $\hat{\lambda}_{c,t}$, we first obtain estimates of F_c by a SUR regression of factors $f_{c,t}$ on lagged common instruments $Z_{c,t-1}$. Then, we obtain $\hat{\Lambda}_c$ through the relation $\nu_c = \text{vec}\left(\Lambda'_c - F'_c\right)$ and $\hat{\lambda}_{c,t} = \hat{\Lambda}_c Z_{c,t-1}$.

A.4 Estimation of the risk premium bias

The bias term for the estimate $\hat{\nu}_c^{WLS}$ of the risk premia is estimated as

$$\hat{B}_{\nu_c} = \hat{Q}_{\beta_3}^{-1} \frac{1}{n_c} \sum_{i=1} \tau_{i,c} J_{b,i} \text{vec}\left(E'_{2,i} \hat{Q}_{x,i,c}^{-1} \hat{S}_{ii,c} \hat{Q}_{x,i,c}^{-1} C_{\hat{\nu}_c} \hat{w}_{i,c}\right), \tag{7}$$

with $J_{b,i} = (\text{vec}(I_{d_{1,i}})' \otimes I_{Kp})(I_{d_{1,i}} \otimes J_{a,i})$.

A.5 Estimation of the risk premium covariance matrix

The covariance matrix for the risk premia estimate $\hat{\nu}_c^{WLS}$ is estimated as

[1] In their additional empirical results, GOS show that a value-weighting scheme does not change point estimate values but can increase confidence intervals due to a precision loss.

$$\hat{\Sigma}_{\nu_c} = \hat{Q}_{\beta_3}^{-1}\hat{S}\hat{Q}_{\beta_3}^{-1}$$

where

$$\hat{S} \quad = \quad \frac{1}{n_c}\sum_{i,j}\frac{\tau_i\tau_j}{\tau_{i,j}}\beta'_{i,c,3}w_{i,c}C'_{\hat{\nu}_c,i}\hat{Q}_{x,i,c}^{-1}\tilde{S}_{ij,c}\hat{Q}_{x,j,c}^{-1}C_{\hat{\nu}_c,j}w'_{j,c}\beta_{j,c,3},$$

.

where $\tau_{i,j,c} = \frac{T_c}{T_{ij,c}}$ and $T_{ij,c} = \sum_t I_{i,c,t}I_{j,c,t}$.

In the above equation, we use a hard thresholded estimator

$$\tilde{S}_{ij,c} = \hat{S}_{ij,c}\mathbf{1}_{||\hat{S}_{ij,c}|| \geq \kappa_{n_c,T_c}},$$

where $\hat{S}_{ij,c} = \frac{1}{T_{ij,c}}\sum_t I_{i,c,t}I_{j,c,t}\hat{\varepsilon}_{i,c,t}\hat{\varepsilon}_{j,c,t}x_{i,c,t}x'_{j,c,t}$, $||\hat{S}_{ij,c}||$ is the Frobenius norm, $\kappa_{n_c,T_c} = M\sqrt{\frac{log(n_c)}{T_c}}$ is a data-dependent threshold, and M is a positive number set by cross-validation (see GOS for details).

Thresholding ensures that the estimator $\hat{S}$ is consistent. Indeed, $\hat{S}$ involves a sum on i and j but is standardized only by n_c (and not n_c^2). Consequently, the usual sample estimator is not consistent. We use in the definition of $\tilde{S}_{ij,c}$ the threshold proposed in Bickel and Levina (2008) and extended by GOS to a random coefficient setting,

A.6 Diagnostic criterion for the factor structure

We compute the T_c-by-T_c matrix Υ

$$\Upsilon = \sum_{i=1}^{n_c}\mathbf{1}_{i,c}^{\chi}\bar{\varepsilon}_{i,c,t}\bar{\varepsilon}'_{i,c,t},$$

where T_c is the number of periods and $\bar{\varepsilon}_{i,c,t}$ is a T_c-by-one vector of standardized residuals $\bar{\varepsilon}_{i,c,t} = \frac{I_{i,c,t}\varepsilon_{i,c,t}}{\sqrt{\frac{1}{T_c}\sum_{t=1}^{T_c}I_{i,c,t}\varepsilon_{i,c,t}^2}}$.

The diagnostic criterion is given by

$$\zeta = eig_{\max}\left(\frac{\Upsilon}{n_cT_c}\right) - g\left(n_c, T_c\right), \tag{8}$$

where $g\left(n_c, T_c\right) = -P\kappa ln\left(\kappa\right)$ is a penalty term with $\kappa = \frac{\left(\sqrt{n_c}+\sqrt{T_c}\right)^2}{n_cT_c}$ and P is a data-driven constant. We use a simulation-based method to select P (see Appendix 7 in GOS2 for details and Monte Carlo results for unbalanced panels).

A.7 Asset pricing test

The test for asset pricing restrictions is based on the weighted sum of squared residuals $\hat{Q}_e = \frac{1}{n_c}\sum_i \hat{e}'_{i,c}\hat{w}_{i,c}\hat{e}_{i,c}$, where $\hat{e}_{i,c} = \hat{\beta}_{i,c,1} - \hat{\beta}_{i,c,3}\hat{\nu}_c^{Unbiased}$. The distribution of the re-centered sum of squared residuals is

$$\tilde{\Sigma}_e^{-1/2} T_c \sqrt{n_c}\left(\hat{Q}_e - \frac{\bar{d}_1}{T_c}\right) \sim N(0,1)$$

where $\tilde{\Sigma}_e = \frac{2}{n_c}\sum_{i,j}\frac{\tau_{i,c}^2\tau_{j,c}^2}{\tau_{i,j,c}^2} Tr\left[\left(C'_{\hat{\nu}_c,i}\hat{Q}_{x,i}^{-1}\tilde{S}_{ij}\hat{Q}_{x,j}^{-1}C_{\hat{\nu}_c,j}\right)\hat{w}_{j,c}\left(C'_{\hat{\nu}_c,j}\hat{Q}_{x,j}^{-1}\tilde{S}_{ji}\hat{Q}_{x,i}^{-1}C_{\hat{\nu}_c,i}\right)\hat{w}_{i,c}\right]$ and $\bar{d}_1 = \frac{1}{n_c}\sum_{i=1}^{n_c} d_{1,i}$.

A.8 Distribution of the risk premium dynamic parameters Λ_c

The parameters for the dynamics of the risk premia, Λ_c, follow a normal distribution

$$\sqrt{T_c}vec[\hat{\Lambda}'_c - \Lambda'_c] \sim N\left(0, \Sigma_{\Lambda_c}\right)$$

where

$$\begin{aligned}
\Sigma_{\Lambda_c} &= \left(\mathbb{I}_K \otimes Q_z^{-1}\right)\Sigma_u\left(\mathbb{I}_K \otimes Q_z^{-1}\right), \\
\Sigma_u &= E\left[u_t u'_t \otimes Z_{c,t-1}Z'_{c,t-1}\right], \\
u_t &= f_{c,t} - F_c Z_{c,t-1}, \\
Q_z &= E\left[Z_{c,t-1}Z'_{c,t-1}\right].
\end{aligned}$$

B Equity data construction

Our objective is to build a database of common stocks traded on major stock exchanges. We examine the pros and cons of using Datastream versus Compustat Global/xpressfeed. Given the longer time series found on CRSP for US stocks, we focus on non-US countries.

Our main conclusions are as follows. Datastream has longer time series for some but not all stocks. However, it contains many errors. Compustat has less data errors, the history of SEDOLs and ISINs, and the type of daily quote which to our knowledge is not available on Datastream (only the current identifiers are available).

The following steps describe how we construct the data for each country. By visual inspection of value- and equal-weighted indexes, we investigate each discrepancy. In some cases, we can confirm a mistake in Datastream (Compustat) by using data from Compustat (Datastream). For example, a spike in the total return index on Datastream is identified and removed by looking at the total return index on Compustat. In other cases, we can not conclude which of the two databases has an error and further check on Bloomberg and/or MSCI.

Given the advantages listed above, we use data from Compustat/xpressfeed in this paper. We describe the filters and error corrections we use for each of the two databases in the following steps. Therefore, this guide can be used for research based on Datastream or Compustat data.

1. **Stock Universe:**

   - Datastream: We retrieve all securities which are classified as equity (*instrument_type = 'Equity'*).
   - Compustat: We retrieve all securities which are classified as common or ordinary shares (*tpci = '0'*).

2. **Major Stock Exchanges:** We keep only stocks listed on a country major stock exchange. We define the major stock exchange as the one with the highest number of listed stocks. In most cases, the choice is obvious. However, we include more than one stock exchanges in a few countries.

3. **Refining the common stock universe:** Securities are misclassified in both databases. We apply the following additional filter on the security name:

   - Datastream: We apply the name and industry filters as in Griffin et al. (2010). We add "BDR" to the list of keywords to remove Brazilean Depositary Receipts. We also use additional keyword filters used by Lee (2011): "AFV" in Belgium due to their preferential tax treatment, "INC.FD." in Canada because they are income trusts, and "RSP" in Italy due to their nonvoting provisions.

- Compustat: We remove non-common stocks based on the presence of the same keywords in their issue description (*dsci*).

4. **Preliminary cleaning of times series:**

   - Compustat: We use only days for which a price (*prccd*) is available with a price code status (*prcstd*) either equal to 3 (high, low and close prices) or 10 (prices as reported). We also include price code status 4 (bid, ask, average/last volume close) for Canadian issues because Compustat historically delivered prices as the average of the bid/ask pricing for U.S. and Canadian issues.
   - Datastream: We use only days for which the unadjusted price (*UP*) is available. Datastream does not provide any indication as to the type of quote it provides. In many cases, total return indexes (*RI*) continue after the price stops quoting. Datastream repeats the last price after a stock stops. For each stock, we verify each day if the rest of the time series is the same price and remove the rest of the time series in such case. This procedure does not capture cases in which a stock stops quoting for a few months and then starts again. In this case, we get a series of zero returns.

     At this stage, indexes built from Datastream have longer time series for many countries compared to Compustat indexes. This is especially the case for some developed countries whose indexes start in the early 1970s whereas all non-North American data on Compustat starts in the early 1980s. However, many unexplained spikes in Datastream time series come from days for which only the price is available. We can match several of these cases to Compustat data and confirm that they correspond to a price standard (*prcstd*) equal to 5 (no price is available, the last price is carried forward). Unfortunately, we cannot match these cases with Compustat data in the pre-1980s period. Therefore, we keep only quotes for which either the volume, low, or high is available as a sign of real market activity. This filter solves many of the initial discrepancies between the two data providers.

5. **Controlling for spikes that are reversed:**

   - Datastream: Following Ince and Porter (2006), we control for extreme daily returns that are reversed the following day. If the total return over two consecutive days is below 50% and any of the two daily total return is above 100%, we remove both daily observations.
   - Compustat: None.

6. **Computing monthly returns:** We build monthly returns by using the last available total return index value during the previous month and the last available value in the current month.

- Datastream: We use the total return index (*RI*). We convert the local total return index to U.S. dollars and keep nine decimals such that monthly returns are not impacted by rounding (using the function *DPL#(X(RI) U$,9)*).
- Compustat: We build total return indexes using prices (*prccd*), adjustment factors (*ajexdi*), quotation units (*qunit*), exchange rates (*exratd*), and total return factors (*trfd*). We follow Shumway (1997) and apply a −30% delisting return when delisting is performance related (using the delisting reason *dlrsni*).

7. **Computing market capitalizations:** We build monthly lagged market capitalizations by using the last available market capitalization during the previous month.

   - Datastream: We use the market value (*MV*) converted to U.S. dollars.
   - Compustat: We build market capitalization by multiplying the number of shares by prices (*prccd*). For non-North American stocks, we use the current number of shares outstanding (*cshoc*). For North-American stocks, we use the last report number of shares outstanding (*cshoi*).

8. **Manual data corrections:** We investigate and identify in Table 1 errors for Compustat data not captured by the filters above.

   In unreported figures available upon request, we plot for each country the returns of the value-weighted and equal-weighted market portfolios as well as the number of stocks over time using both databases.

gvkey/iid	Error
202192/01W, 203051/01W, 207206/01W, 208514/01W	In January 1992 in Argentina, there are four stocks for which the transition from the old currency code ARA to ARS creates 10,000+ returns. We remove them for this month.
203579/01W, 205247/01W	Before January 1992 in Argentina, these two stocks' USD market capitalization are off by a factor 10. We multiply the market capitalization by 0.1.
029178/01W	This Argentinean stock's market cap is too large and erratic, and there are some holes. Its data on Datastream starts on January 1992. We start in October 1990 after the last hole when the market capitalization is not erratic.
208536/01W	The adjustment factor *ajexdi* does not adjust for the 0.0513-to-1 stock split on May 20th, 2015. We remove the stock for this month.
030581/01W	Before February 1992, this stock in Brazil has extreme market values.
All stocks in Brazil	In January 1989, the 1-to-1,000 change from the Cruzado to the Cruzado novo is not reflected in Compustat's exchange rate table (nor is the one in 1986). We divide returns by 1,000.
206477/01W	There is an error in the adjustment factor (*ajexdi*) from 01/09/2007 to 20/3/2007, it should be 1 instead of 10, verified on Bloomberg.
208194/02W 203187/01W 229956/02W 208200/01W 203462/01W 203682/01W 208603/01W 208366/01W 209409/01W	Spike for these Chinese stocks in March and June 1993. Spike for 203187/01W in June 1993 is confirmed with Bloomberg (but return of 700% happens in July). Datastream show missing infrequent returns for these months. We check all large returns on June 1993 with Bloomberg and we can confirm all but one. We multiply the return in March by 10 and divide by 10 in June.
213573/01W	In February 2002 in Estonia, we replace the 25th return with the 21st, Datastream ends on the 21st. We set R = 0.0111301630700127 / 0.0645498918825071 - 1.
103255/01W, 210759/01W, 240641/01W	There are errors caused by the change of currency to the Euro for these three European stocks. We remove them for January 1999.
All stocks in Iceland	For Iceland, the currency plummets on Oct 8th, 2008 and doubles on February 2nd, 2009. We cannot find this plunge on Bloomberg nor on Yahoo. We use Datastream exchange rates, namely, FX rate 0.009452, 0.008440, 0.006994, 0.008246, 0.008773, and 0.008778 for the month of September 2008 through February 2009.
200503/01W	Spike in price creates a return of 15. This Peruvian stock is not on Datastream and it starts in 1996 on Bloomberg. We remove it for December 1992.
All Peruvian stocks	In January 1992, the 1,000,000-to-1 change described below (from Wikipedia) is not reflected on CSXF. "Because of the bad state of economy and hyperinflation in the late 1980s the government was forced to abandon the inti and introduce the sol as the country's new currency. The currency was put into use on July 1, 1991 (by Law No. 25,295) to replace the inti at a rate of 1 sol to 1,000,000 intis. Coins denominated in the new unit were introduced on October 1, 1991 and the first banknotes on November 13, 1991. Hitherto, the sol has retained a low inflation rate of 1.5%, the lowest inflation rate ever in both Latin and South America. Since the new currency was put into effect, it has managed to maintain a stable exchange rate between 2.2 and 3.66 per United States dollar." We divide returns by 1,000,000.

201673/01W	In July 1998, this New Zealand stock has the same price as on Datastream, but its adjustment factor (*ajexdi*) and total return factor (*trfd*) create a huge difference compared to Datastream. We remove it for this month.
206463/03W	Moscow City Telephone Network Co has random 1000x spikes in the price time series, it would take too many corrections to solve the problem. We remove the complete time series.
284439/01W	In January 2005, there is an error in the adjustment factor (*ajexdi*) when the currency changed. Other stocks' prices (*prccd*) and *ajexdi* adjust. This stock *prccd* adjusts, but not its *ajexdi*. We remove it for this month.
217719/01W	In February and March 1994, there is an error for this Colombian stock (verified with Datastream) and remove it for those two months.
185208/01C	This Canadian stock is delisted on January 1^{st} 2017, there is a spike in the price on December 30^{th}, 2016, and the time series ends on December 2^{nd}, 2016, on Bloomberg. We remove it for December 2016. CSXF is also missing the total return adjustment for the 100-to-1 conversion on November 1^{st}, 2013, which creates a 100+ % return. We remove it for November 2013.
202022/01W	This Chilean stock has erratic and infrequent quotes before January 2004. There are price spikes on days with unavailable volumes, but classified as "prices as reported" (*prcstd*=10). There are no quotes on these days on Bloomberg. We remove infrequent returns before January 2004.
149822/01C	The number of shares outstanding (*cshoc*) is off by a factor 100 for the last two days of June 2004. We then correct the number of shares.

Table 1 We report in this table the manual data corrections to data on Compustat/xpressfeed.

C Which instruments are important for time-varying factor exposures in regional models?

In this section, we provide the median coefficients for each factor and each instrument for the mixed regional four-factor model in Table 2 and for the mixed regional *q*-factor model in Table 3. These tables have the same structure as Tables 4-5 in the main text.

Table 2 Which instruments drive time-variations in factor exposures in the mixed regional four-factor model?

Factor	Region	Constant	Country dividend yield	Size	Value	Momentum
		(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*
Market	Developed Markets	0.93	0.01	−0.23	−0.03	0.08
		(100.00)	(81.09)	(78.12)	(81.51)	(90.05)
	Emerging Markets	0.94	0.01	−0.03	0.20	0.08
		(100.00)	(84.07)	(79.27)	(80.63)	(88.35)
Size	Developed Markets	0.73	0.04	−1.23	0.21	−0.03
		(100.00)	(80.75)	(79.14)	(82.79)	(90.76)
	Emerging Markets	0.56	−0.03	−0.83	0.15	0.01
		(100.00)	(84.75)	(79.26)	(81.62)	(90.99)
Value	Developed Markets	0.03	−0.07	−0.32	0.75	−0.04
		(100.00)	(70.58)	(73.13)	(77.29)	(82.78)
	Emerging Markets	0.03	0.00	−0.17	0.52	−0.08
		(100.00)	(83.58)	(77.41)	(80.13)	(87.96)
Momentum	Developed Markets	0.06	−0.01	−0.30	−0.15	0.39
		(100.00)	(74.47)	(72.33)	(76.64)	(84.49)
	Emerging Markets	0.02	−0.03	0.15	0.09	0.25
		(100.00)	(82.71)	(79.00)	(82.64)	(90.46)
Excess country market	Developed Markets	0.85	0.05	0.32	0.13	−0.02
		(100.00)	(82.76)	(79.46)	(83.93)	(91.21)
	Emerging Markets	1.01	−0.02	0.03	0.18	0.01
		(100.00)	(85.37)	(79.95)	(81.66)	(90.07)
Median time-series R^2 (%)	Developed Markets	25.95				
	Emerging Markets	38.14				

We report the median coefficient value for each factor-instrument interaction in the time-series regressions for the mixed regional four-factor model. For each factor in the first column, we report the median coefficient for each instrument in columns (i) to (v) across all stocks in developed markets and across all stocks in emerging markets. Below each coefficient median, we report in parentheses the proportion (in %) of stocks for which the regressor is selected by our methodology. Finally, we report in the last rows the median time-series regression R^2.

Table 3 Which instruments drive time-variations in factor exposures in the mixed regional q-factor model?

Factor	Region	Constant	Country dividend yield	Size	Value	Momentum
		(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*
Market	Developed Markets	0.89	0.00	−0.12	−0.03	−0.00
		(100.00)	(82.38)	(83.11)	(83.68)	(88.04)
	Emerging Markets	0.94	−0.01	−0.23	0.00	0.00
		(100.00)	(87.04)	(84.73)	(86.09)	(88.74)
Size	Developed Markets	0.67	−0.02	−1.75	0.03	0.06
		(100.00)	(83.23)	(83.90)	(85.08)	(89.13)
	Emerging Markets	0.52	−0.00	−1.07	−0.04	0.03
		(100.00)	(88.94)	(84.68)	(88.08)	(91.05)
Profitability	Developed Markets	−0.01	−0.05	0.30	0.32	0.03
		(100.00)	(82.02)	(83.76)	(84.84)	(88.84)
	Emerging Markets	0.05	−0.03	0.48	0.29	−0.03
		(100.00)	(90.09)	(86.23)	(89.18)	(91.91)
Investment	Developed Markets	0.06	−0.09	−1.04	0.07	−0.37
		(100.00)	(82.01)	(82.46)	(84.03)	(89.03)
	Emerging Markets	0.07	0.01	−0.78	−0.13	−0.36
		(100.00)	(88.19)	(85.36)	(88.39)	(91.66)
Excess country market	Developed Markets	0.75	0.04	0.01	−0.02	0.01
		(100.00)	(83.42)	(83.65)	(85.49)	(89.71)
	Emerging Markets	1.01	−0.00	−0.10	−0.01	0.01
		(100.00)	(88.80)	(85.12)	(87.49)	(90.31)
Median time-series R^2 (%)	Developed Markets	24.80				
	Emerging Markets	36.61				

We report the median coefficient value for each factor-instrument interaction in the time-series regressions for the mixed regional q-factor model. For each factor in the first column, we report the median coefficient for each instrument in columns (i) to (v) across all stocks in developed markets and across all stocks in emerging markets. Below each coefficient median, we report in parentheses the proportion (in %) of stocks for which the regressor is selected by our methodology. Finally, we report in the last rows the median time-series regression R^2.

References

Andrews, Donald W. K., 2005, Cross-section regression with common shocks, *Econometrica* 73, 1551–1585.

Belsley, David A., Edwin Kuh, and Roy E. Welsch, 2004, *Regression diagnostics - Identifying influential data and sources of collinearity* (John Wiley & Sons, New York).

Bickel, Peter J., and Elizaveta Levina, 2008, Covariance regularization by thresholding, *The Annals of Statistics* 36, 2577–2604.

Connor, Gregory, and Robert A. Korajczyk, 1987, Estimating pervasive economic factors with missing observations, *Working Paper No. 34, Department of Finance, Northwestern University* .

Gagliardini, Patrick, Elisa Ossola, and Olivier Scaillet, 2016, Time-varying risk premium in large cross-sectional equity datasets, *Econometrica* 84, 985–1046.

Greene, William H., 2008, *Econometric Analysis, 6th ed.* (Prentice Hall).

Griffin, John M., Patrick J. Kelly, and Federico Nardari, 2010, Do market efficiency measures yield correct inferences? a comparison of developed and emerging markets, *Review of Financial Studies* 23, 3225–3277.

Hsiao, Cheng, 2003, *Analysis of Panel Data* (Econometric Society Monographs, 2nd edition, Cambridge University Press).

Ince, Ozgur S., and R. Burt Porter, 2006, Individual equity return data from Thompson Datastream: handle with care!, *Journal of Financial Research* 29, 463–479.

Lee, Kuan-Hui, 2011, The world price of liquidity risk, *Journal of Financial Economics* 99, 136–161.

Shumway, Tyler, 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.