The mosaic permutation test: an exact and nonparametric goodness-of-fit test for factor models

Asher Spector * Rina Foygel Barber † Trevor Hastie * Ronald N. Kahn ‡
Emmanuel Candès *§

April 24, 2024

Abstract

Financial firms often rely on factor models to explain correlations among asset returns. These models are important for managing risk, for example by modeling the probability that many assets will simultaneously lose value. Yet after major events, e.g., COVID-19, analysts may reassess whether existing models continue to fit well: specifically, after accounting for the factor exposures, are the residuals of the asset returns independent? With this motivation, we introduce the mosaic permutation test, a nonparametric goodness-of-fit test for preexisting factor models. Our method allows analysts to use nearly any machine learning technique to detect model violations while provably controlling the false positive rate, i.e., the probability of rejecting a well-fitting model. Notably, this result does not rely on asymptotic approximations and makes no parametric assumptions. This property helps prevent analysts from unnecessarily rebuilding accurate models, which can waste resources and increase risk. We illustrate our methodology by applying it to the Blackrock Fundamental Equity Risk (BFRE) model. Using the mosaic permutation test, we find that the BFRE model generally explains the most significant correlations among assets. However, we find evidence of unexplained correlations among certain real estate stocks, and we show that adding new factors improves model fit. We implement our methods in the python package `mosaicperm`.

1 Introduction

1.1 Motivation and problem statement

Factor models are perhaps the most common statistical tool used to manage risk in economics and finance (Grinold and Kahn, 1994). Indeed, analysts routinely use factor models to model the correlations between asset returns, allowing one to estimate the probability that many assets in a portfolio will simultaneously lose value. Yet as conditions change, analysts must assess whether established models continue to be reliable. As an illustrative example, this paper analyzes the BlackRock Fundamental Equity Risk (BFRE) model, one of many commercially available risk models used in industry. In particular, Section 4 asks whether the BFRE model adequately explains correlations among US stock returns two months after the COVID-19 pandemic began. Correctly answering such questions

*Department of Statistics, Stanford University
†Department of Statistics, University of Chicago
‡BlackRock, Systematic Investment Research
§Department of Mathematics, Stanford University

is essential: on the one hand, needlessly rebuilding an established factor model may waste resources and ultimately increase risk, but on the other hand, it is important to quickly detect inadequacies in existing models.

This article develops statistical methods to test the goodness-of-fit of existing factor models. Formally, at times $t = 1, \dots, T$, suppose we observe returns $Y_t \in \mathbb{R}^p$ for p assets which we believe follow the factor model

$$Y_t = L_t X_t + \epsilon_t, \tag{1.1}$$

where L_t, X_t and ϵ_t are defined below:

- $X_t \in \mathbb{R}^k$ denotes the returns of $k \ll p$ underlying factors which drive correlation among the assets. We assume the factor returns X_t are not observed.
- $L_t \in \mathbb{R}^{p \times k}$ are factor "loadings" or exposures, i.e., $[L_t]_{j\ell}$ measures the exposure of the jth asset to the ℓth factor at time t. We treat L_t as a deterministic matrix which is known at time t (see below).
- $\epsilon_t \in \mathbb{R}^p$ denotes the idiosyncratic returns of the p assets which cannot be explained by the factors. We also refer to ϵ_t as the "residuals."

This paper primarily analyzes *fundamental risk models* like the BFRE model, where the exposures L_t are based on market fundamentals such as industry membership and accounting data (Grinold and Kahn, 1994). For example, $[L_t]_{j\ell} \in \{0, 1\}$ might indicate whether stock j is in the ℓth industry. Unlike factor models commonly used in, e.g., psychology, this means that the exposures L_t are known at time t, although the factor returns X_t are not observed and must be estimated, typically using cross-sectional regressions. Naturally, other risk models exist, including (i) *macrofactor risk models*, where X_t denotes observed macroeconomic time series data and L_t is unknown, and (ii) *purely statistical models*, where both X_t and L_t are estimated. Such models are beyond the scope of this paper.[1]

To test if (1.1) is an adequate model, let $\epsilon_{\cdot,j} \coloneqq (\epsilon_{1,j}, \dots, \epsilon_{T,j}) \in \mathbb{R}^T$ denote the time series of residuals for the jth asset. We will test the null that the residual processes are independent across assets:

$$\mathcal{H}_0 : \epsilon_{\cdot,1}, \epsilon_{\cdot,2}, \dots, \epsilon_{\cdot,p} \in \mathbb{R}^T \text{ are jointly independent.} \tag{1.2}$$

Note that $\mathcal{H}_0$ allows there to be temporal dependence among the residuals of the jth asset, but it requires all residuals of the jth asset to be independent of all other residuals. If $\mathcal{H}_0$ holds, we say that (1.1) accurately models the correlations among asset returns. We emphasize that we seek to test whether $\mathcal{H}_0$ holds for a *fixed* choice of exposures L_t, motivated by the fact that many existing financial risk models routinely publish exposure matrices L_t, including MSCI Barra models and the BFRE model (Rosenberg and Marathe, 1976; Bender and Nielsen, 2012).[2] In contrast, many previous works test whether $\mathcal{H}_0$ holds for some *unknown* choice of $L_t \in \mathbb{R}^{p \times k}$ or estimate the number of factors k (see Section 1.4). These problems have other applications, but they do not accomplish our goal: if one's own risk model is highly misspecified, it is of little comfort to know that some unknown choice of factor loadings is well-specified.

We argue that a good test of $\mathcal{H}_0$ should rigorously control false positives, i.e., it should reject $\mathcal{H}_0$ with probability at most α whenever $\mathcal{H}_0$ actually holds. This is important for several reasons. First, it is important for risk management: in times of volatility, needlessly doubting a well-fitting risk model could be just as harmful as relying on a misspecified one. Second, large financial firms may constantly stress-test their risk models. Without rigorous false positive control, they may discard and rebuild

[1] It is possible to extend the methods in this paper to the case where X_t is observed and L_t is not. However, it requires rather different statistical techniques, so we defer this extension to a companion paper (Spector et al., 2024).

[2] Naturally, our methods also apply if one selects exposures using historical data and tests the selected model's goodness-of-fit on fresh data.

many well-fitting models for no reason, consuming a great deal of resources and possibly reducing model quality in the long run. Lastly, rigorous hypothesis tests may be helpful during the process of constructing the exposures L_t, because they quantify evidence against different candidate models.

These arguments are not new, and the problem of testing the goodness-of-fit of a factor model dates back to the origins of the field of statistics (Spearman, 1904; Roy, 1953; Bartlett, 1954; Box and Andersen, 1955; Lawley, 1956; Horn, 1965). These seminal works established that if (i) the idiosyncratic returns ϵ_t are Gaussian or (ii) the number of assets p is held constant as the number of timepoints T diverges, one can perform (asymptotically) valid hypothesis tests using the generalized likelihood ratio (GLR) test—see Anderson (2009) for a review. Alternatively, another classical approach to uncertainty quantification would be to apply the bootstrap (Efron, 1979) or the block bootstrap (e.g., Kunsch, 1989; Romano and Wolf, 2006).

However, these classical techniques may not be suited to modern financial datasets, for several reasons. First, modern applications are typically high-dimensional, meaning that the number of assets p is comparable to or much larger than T. Indeed, to quickly detect large violations of $\mathcal{H}_0$, we might analyze datasets with $p \geq 2000$ assets and $T \approx 50$ datapoints. In these settings, classical theory for likelihood ratios and bootstrap methods will be generally inaccurate (e.g., Karoui and Purdom, 2018; Sur and Candès, 2019). Even modern asymptotic results for high-dimensional factor models (e.g., Bai, 2003) may be inaccurate when T is small (see Section 1.4 for review). Second, parametric assumptions are not appropriate, since real data may exhibit features that are not captured by the model, such as heavy tails or heteroskedasticity. Lastly, even if we could apply classical likelihood-based theory, we might not want to use the GLR test statistic, because we would prefer to use regularization or other machine learning techniques to increase power. Thus, in this paper, we ask: can we develop finite-sample valid tests of $\mathcal{H}_0$ under no parametric assumptions? Additionally, can we do so in a way that leverages prior information and black-box machine learning techniques to increase power while retaining false positive control?

1.2 A motivating application to the BlackRock Fundamental Equity Risk model

As a motivating example, we analyze the BlackRock Fundamental Equity Risk (BFRE) model, a factor model which publishes weekly exposure matrices L_t for $p \geq 2000$ US stocks and $k \approx 65$ factors spanning all sectors of the US economy. We ask: two months after the COVID-19 pandemic began, does the BFRE model adequately explain the correlations among US stock returns?

One reasonable way to answer this question might be to track, for each asset, the maximum absolute correlation between its residuals and those of another asset. To make this idea precise, we:

1. Run cross-sectional ordinary least squares (OLS) regressions to estimate the residuals.[1] I.e., let $H_t^{\text{ols}} := I_p - L_t(L_t^\top L_t)^{-1}L_t^\top$ denote the standard OLS projection matrix, and define:

$$\hat{\epsilon}_t^{\text{ols}} := H_t^{\text{ols}} Y_t = H_t^{\text{ols}}(L_t X_t + \epsilon_t) = H_t^{\text{ols}}\epsilon_t. \tag{1.3}$$

2. Compute the empirical correlation matrix $\hat{C} \in \mathbb{R}^{p\times p}$ of the last 350 estimated residuals $\{\hat{\epsilon}_s^{\text{ols}}\}_{s=t-350}^{t}$. (The choice of window size is somewhat arbitrary, but Appendix B.1 confirms that we obtain similar results using many different window sizes.)

3. For each asset $j \in [p]$, define $\widehat{\text{MaxCorr}}_j := \max_{j'\neq j} |\hat{C}_{j,j'}|$ as the maximum estimated absolute correlation between asset j and another asset.

[1] As noted in Section 1, estimating ϵ via cross-sectional regressions is standard for factor models with known L_t.

4. Finally, let $S_t^{\text{ols}} = \frac{1}{p}\sum_{j=1}^{p} \widehat{\text{MaxCorr}}_j$ denote the *mean maximum (absolute) correlation* (MMC) over all assets at time t, which we use as an aggregate measure of model fit. Indeed, Figure 1 plots S_t^{ols} biweekly for three sectors between 2018 and 2023.

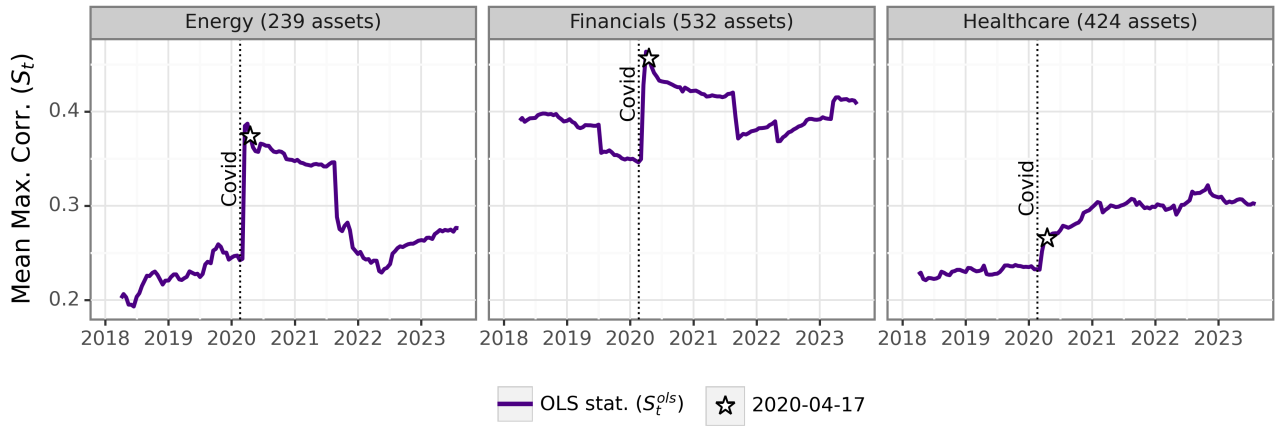


Figure 1: For three industries, this figure plots biweekly values of $S_t^{\text{ols}} = \frac{1}{p}\sum_{j=1}^{p} \widehat{\text{MaxCorr}}_j$, where $\widehat{\text{MaxCorr}}_j = \max_{j' \neq j} |\hat{C}_{j,j'}|$ and $\hat{C}$ is the empirical correlation matrix of the estimated idiosyncratic returns $\{\hat{\epsilon}_s^{\text{ols}}\}_{s-350}^{t}$ from the last 350 days. Interpreting this plot is challenging because it is not obvious what curve one would expect to see even if the factor model fits perfectly. In general, S_t^{ols} is neither mean-zero nor stationary under $\mathcal{H}_0$. Indeed, the large jumps between February and April 2020 coincide with large increases in the variance of ϵ_t.

Producing plots like Figure 1 is relatively easy—however, *interpreting* these plots is hard, because it is not obvious what types of fluctuations we would see even if the null were true. For example, even when ϵ_t has independent components, $\text{Cov}(\hat{\epsilon}_t^{\text{ols}}) = H_t^{\text{ols}} \text{Cov}(\epsilon_t) H_t^{\text{ols}}$ is not diagonal, so even under the null, we would expect to see some correlations among $\hat{\epsilon}_t^{\text{ols}}$. Furthermore, even under $\mathcal{H}_0$, the law of S_t^{ols} should change over time, because (a) the projection matrices H_t^{ols} change over time due to the changing exposures and (b) the variances of the idiosyncratic returns $\text{Cov}(\epsilon_t^{\text{ols}})$ change over time—for example, after the COVID pandemic began, the variance of each idiosyncratic return increased dramatically.

As a result, it is hard to know how to interpret Figure 1. Certainly in all three sectors, the value of S_t jumps after February 2020—but is this jump consistent with the existing factor model? For example, on April 17th, 2020, we observe absolute correlations of $0.38, 0.43$, and 0.27 in the energy, financial, and healthcare sectors, respectively. Should we be concerned by these numbers? Our work aims to provide statistical tools to answer these questions, not only for this particular test statistic, but for many measures of the goodness-of-fit of the model.

1.3 Contribution

Our paper introduces an exact and nonparametric permutation test of $\mathcal{H}_0$. The key idea is to introduce a new estimator $\hat{\epsilon} \in \mathbb{R}^{T \times p}$ of the residuals which exactly preserves some of the independence properties of the true residuals ϵ. To construct $\hat{\epsilon}$, we split the data matrix $\mathbf{Y}$ into rectangular tiles (along both axes) and separately estimate the residuals in each tile, yielding an estimate $\hat{\epsilon} \in \mathbb{R}^{T \times p}$ of the residuals. Given a test statistic $S(\hat{\epsilon})$ quantifying the correlations among the columns of $\hat{\epsilon}$, we can compute the law of $S(\hat{\epsilon})$ under the null by permuting the rows of each tile of $\hat{\epsilon}$. This idea is illustrated in Figure 2. We refer to this as a "mosaic permutation test" because the separation of the data into tiles is reminiscent of a mosaic.

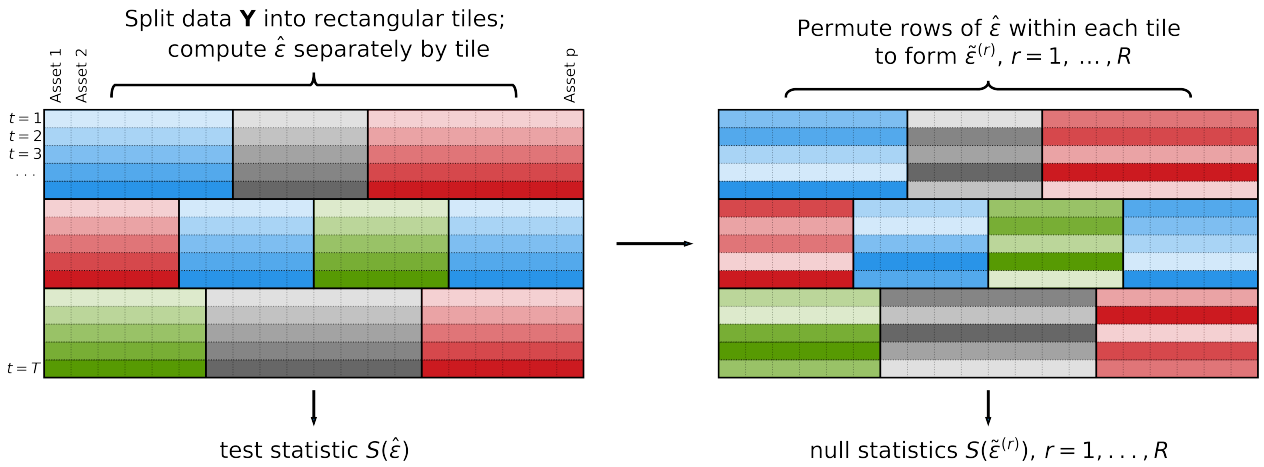


Figure 2: This figure summarizes the main methodology of the mosaic permutation test. Above, rows represent different observations $t = 1, \ldots, T$, columns represent different assets, and in the second matrix the shadings are permuted within each rectangle to illustrate the permutations within tiles.

The mosaic permutation test has three key properties.

1. Exact and nonparametric false positive control. The test yields an exact p-value in finite samples under only the assumption that the idiosyncratic returns for each asset are independently and identically distributed (i.i.d.), or more generally, that they are locally exchangeable (defined in Section 3). In particular, we make no assumptions about the marginal distributions of $\{\epsilon_t\}_{t=1}^T$ and $\{X_t\}_{t=1}^T$, allowing them to be arbitrarily heavy-tailed and heteroskedastic. Furthermore, to allow for changing market conditions, our results allow the factor returns $\{X_t\}_{t=1}^T$ to be arbitrarily non-stationary and the idiosyncratic returns $\{\epsilon_t\}_{t=1}^T$ to be non-stationary across tiles.

To illustrate this contribution, we conduct semisynthetic simulations using the exposure matrix L_t from the BFRE model for financial stocks on April 17th, 2020. For simplicity, we generate new data **Y** from Eq. 1.1 after sampling the residuals and factor returns as i.i.d. standard Gaussians, with $T = 350$ observations. We use the same test statistic as in Section 1.2, after computing $\hat{\epsilon}^{\text{ols}}$ using cross-sectional regressions. Figure 3 shows that even in this simple Gaussian setting, naive bootstrap and permutation testing methods both yield essentially a 100% false positive rate (we review these methods and intuition for their failure in Section 2). In contrast, the mosaic permutation test has provable validity in finite samples.

2. Power and flexibility. The test allows analysts to use nearly any test statistic to quantify evidence against the null while retaining provable false positive control. For example, it allows analysts to use regularized estimates of the covariance matrix of the idiosyncratic returns $\text{Cov}(\epsilon_t)$, for example, via a graphical lasso (Friedman et al., 2007), and it also permits the use of cross-validation to choose the strength of the regularization. The only restriction is that the test statistic must be a function of the mosaic estimator $\hat{\epsilon}$ of the residuals, instead of a function of (e.g.) a naive OLS estimator $\hat{\epsilon}^{\text{ols}}$ of the residuals. This is the price we pay for rigorous uncertainty quantification, since in practice we cannot easily check the significance of an arbitrary test statistic $S(\hat{\epsilon}^{\text{ols}})$.

That said, in semisynthetic simulations based on US stock data, we find that using the mosaic estimator $\hat{\epsilon}$ does not substantially reduce power compared to an "oracle" test which uses $\hat{\epsilon}^{\text{ols}}$ (see Figure 10).

3. Interpretable test statistics. Our primary methodological contribution is to develop finite-sample tests of $\mathcal{H}_0$. However, once one has rejected $\mathcal{H}_0$, it may also be of interest to learn how to *improve* the factor model, and our empirical analysis (Section 4.3) develops test statistics that can

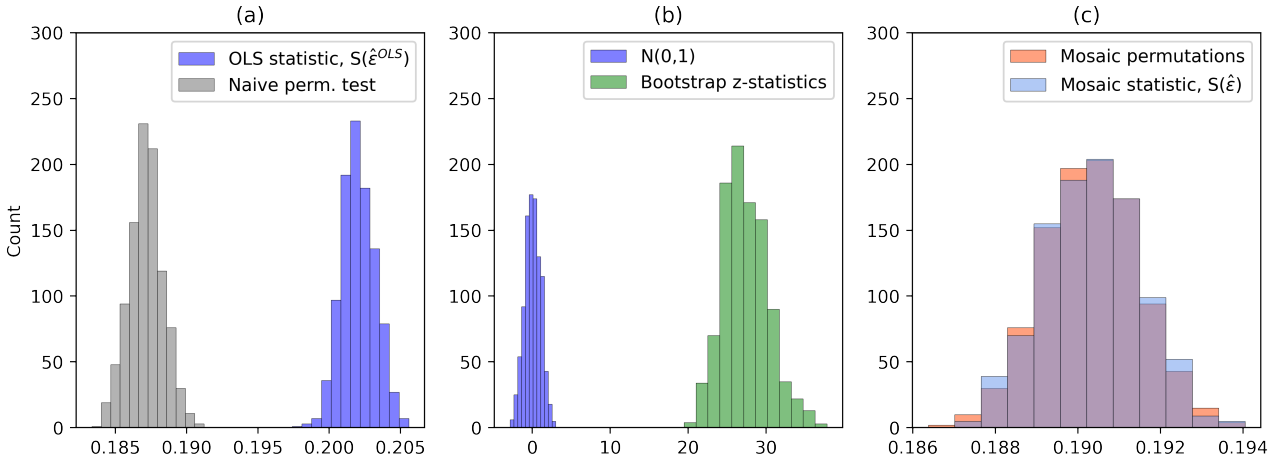


Figure 3: Semisynthetic simulation with $X_{tk}, \epsilon_{tj} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and the exposures L are taken from the BFRE model for the financial sector on April 17th, 2020. Note $T = 350$, $p = 532$, $k = 27$ and we use the same test statistic as in Figure 1. Panel 2(a) shows that a naive residual permutation test (discussed in Section 2.1) inaccurately simulates the null distribution of the test statistic $S(\hat{\epsilon}^{\text{ols}})$—in fact, the true null distribution and the estimated one do not overlap. Panel 2(b) shows that naive bootstrap Z-statistics (discussed in Section 2.2) are not approximately mean zero, nor do they have unit variance. All p-values based on these two naive methods are numerically indistinguishable from zero, leading to an empirical false positive rate of 100%. In contrast, the mosaic permutation test uses a different "mosaic" estimator of the residuals $\hat{\epsilon}$. Using $\hat{\epsilon}$ in place of $\hat{\epsilon}^{\text{ols}}$ allows us to use a permutation method to accurately simulate the law of $S(\hat{\epsilon})$ under the null, as shown in Panel 2(c)—see Section 3 for details.

help answer this question. For example, consider a setting where at least one factor exposure is missing from L_t:

$$Y_t = L_t X_t + v Z_t + \epsilon_t, \tag{1.4}$$

where $v \in \mathbb{R}^p$ denotes the missing factor exposures and $Z_t \in \mathbb{R}$ denotes a missing factor. Motivated by this setting, Section 4.3 introduces practical test statistics which (i) adaptively estimate the sparsity of any missing factor exposures v and (ii) are designed to diagnose when an estimate $\hat{v}$ of missing factor exposures truly improves the model fit.

Empirical application: To illustrate our methods, we test the goodness-of-fit of the BlackRock Fundamental Equity Risk (BFRE) model, from 2018 through 2023. Our analysis distinguishes between *persistent* factors, which retain their explanatory power over long periods of time, and *transient* (non-persistent) factors. (This difference is relevant since commercial models typically focus on including persistent factors but not necessarily transitory factors.) We report three overall findings:

1. The BFRE model appears to fit well in most sectors of the economy, where we are unable to substantially improve the model fit.
2. However, we find evidence of statistically significant unexplained correlations (i) among real estate stocks and (ii) among healthcare stocks post-COVID. The unexplained correlations among healthcare stocks appear to be *transient*, as incorporating them does not consistently improve the model. However, we show that adding an additional factor to account for extra correlations among real estate stocks *persistently* improves the model fit.
3. In contrast, removing existing BFRE factors from the model leads to much stronger evidence against the null.

We present our findings in more detail in Section 4—however, for illustration, Figure 4 now shows the results after applying the mosaic permutation test to the test statistics from Figure 1. To be concrete,

as of April 17th, 2020, we do not find statistically significant evidence against $\mathcal{H}_0$ in the energy sector, but we do find evidence of unexplained correlations in the financial and healthcare sectors. This result is not obvious before one tests for statistical significance, as Figure 4 confirms the intuition from Section 1.2 that the significance threshold for S_t is not constant over time.

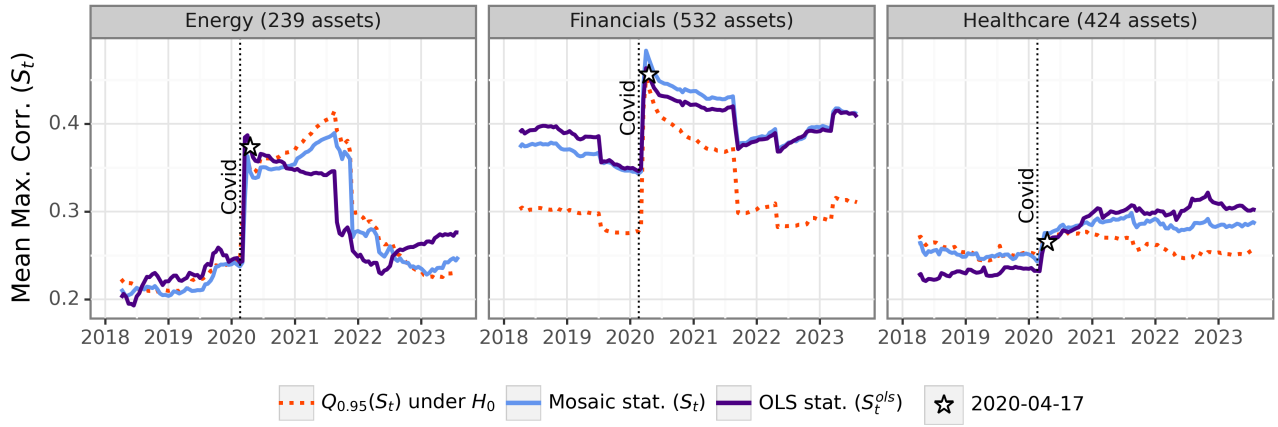


Figure 4: This plot is identical to Figure 1 with two additions. First, we compute a mosaic version (S_t) of the OLS test statistic (S_t^{ols}) by replacing the OLS estimates $\hat{\epsilon}_t^{\text{ols}}$ of the residuals with mosaic estimates $\hat{\epsilon}_t$—see Sections 3 and 4 for details. Second, this allows us to check the statistical significance of S_t using mosaic permutations. The dotted orange line displays the 95% quantile of S_t under $\mathcal{H}_0$—in other words, S_t is statistically significant at time t if the blue line lies above the orange line. Note that to better visualize the correlation between S_t^{ols} and S_t, in this plot only, we shift the test statistic S_t up by a small constant shift (0.06, 0.03 and 0.06 respectively) so that average plotted values of S_t and S_t^{ols} are equal—however, this constant shift provably does not affect the mosaic p-value since $Q_{0.95}(S_t)$ shifts upwards by the same amount.

1.4 Additional related literature

Beyond the classical methods mentioned in Section 1.1, our work contributes to a wide literature on inference for factor models, including inference on the factor exposures L in the high-dimensional setting where p grows with T (Bai, 2003), inference on the number of factors k (e.g., Onatski, 2009; Alessi et al., 2010; Owen and Wang, 2016; Dobriban and Owen, 2018; Dobriban, 2020), tests for changepoints in the factor loadings (e.g., Breitung and Eickmeier, 2011; Bai et al., 2022), tests for whether observed proxies of the factor returns X_t are good proxies (Bai and Ng, 2006), bootstrap methods to debias OLS estimates of L (Gonçalves and Perron, 2020), and more—see Bai and Wang (2016) for a review. Many of these techniques leverage key results from random matrix theory (Johnstone, 2001; Paul, 2007; Bai and Silverstein, 2010), sometimes in combination with permutation-based methods (Buja and Eyuboglu, 1992).

However, our work differs from this existing literature in three respects. First, we solve a different problem—motivated by real financial applications, we seek to test the goodness-of-fit of a model with known exposures $\{L_t\}_{t\in[T]}$. In contrast, existing works treat the exposures as unknown nuisance parameters, and we are not aware of any existing works that explicitly test $\mathcal{H}_0$ for a known choice of exposures. Second, even for the problems they solve, existing works only provide *asymptotic* control of the false positive rate under technical assumptions controlling the heavy-tailedness of the data and the stationarity of both the factors and the idiosyncratic returns (see Bai and Ng (2008); Bai and Wang (2016) for longer reviews of these assumptions). It is not clear that these methods can satisfactorily control false positives when T is small (for example, $T \leq 50$) or when the technical assumptions are

violated. In contrast, our method exactly controls false positives in finite samples assuming only a local exchangeability condition (see Section 3). Lastly, the vast majority of existing works require the analyst to use specific test statistics, such as likelihood ratios or the eigenvalues of the empirical covariance matrix (e.g., Bai, 2003; Onatski, 2009; Alessi et al., 2010; Breitung and Eickmeier, 2011; Bai et al., 2022). In contrast, our work allows analysts to use regularization and black-box machine learning techniques to quantify evidence against the null.

Finally, our work contributes to a growing literature on exact finite-sample permutation tests for linear models (Lei and Bickel, 2020; Wen et al., 2023; D'Haultfœuille and Tuvaandorj, 2023; Guan, 2023). However, these tests are not designed to apply to factor models, and they would apply only if all idiosyncratic returns have the same distribution, which is not realistic, since (e.g.) the variance of the idiosyncratic returns typically varies substantially across assets. In contrast, our theory allows the distribution of the idiosyncratic returns to vary arbitrarily across assets.

1.5 Notation

For $n \in \mathbb{N}$, define $[n] \coloneqq \{1, \ldots, n\}$. For any $A \in \mathbb{R}^{n_1 \times n_2}$, $A_i \in \mathbb{R}^{n_2}$ denotes the ith row of A, and $A_{\cdot,j} \in \mathbb{R}^{n_1}$ denotes the jth column of A. For subsets $I \subset [n_1], J \subset [n_2]$, $A_I \in \mathbb{R}^{|I| \times n_2}$ denotes the submatrix formed by the rows in I, $A_{\cdot,J} \in \mathbb{R}^{n_1 \times |J|}$ denotes the submatrix formed by the columns in J, and $A_{I,J} \in \mathbb{R}^{|I| \times |J|}$ denotes the submatrix formed by the rows in I and the columns in J. We let $\epsilon \coloneqq \begin{bmatrix} \epsilon_1 & \ldots & \epsilon_T \end{bmatrix}^\top \in \mathbb{R}^{T \times p}$ denote the matrix of residuals and $\mathbf{Y} \in \mathbb{R}^{T \times p}$ denotes the observed returns. Thus, $\epsilon_t \in \mathbb{R}^p$ denotes the vector of all p assets' residuals at time t, whereas $\epsilon_{\cdot,j} \in \mathbb{R}^T$ denotes the time series of residuals for asset j. We let $\hat{\epsilon}^{\text{ols}} \in \mathbb{R}^{T \times p}$ denote estimates of the residuals formed using cross-sectional OLS regressions, as defined in Eq. 1.3. $\hat{\epsilon} \in \mathbb{R}^{T \times p}$ denotes the proposed mosaic estimates of the residuals, as introduced in Section 3.

2 Performance of default bootstrap and permutation methods

To review from Section 1, the problem statement is to test the following factor model:

$$Y_t = L_t X_t + \epsilon_t \text{ for } t = 1, \ldots, T, \tag{2.1}$$

for outcomes $Y_t \in \mathbb{R}^p$, fixed and known exposures $L_t \in \mathbb{R}^{p \times k}$, unobserved factor returns $X_t \in \mathbb{R}^k$ and residuals $\epsilon_t \in \mathbb{R}^p$. We seek to test the null hypothesis $\mathcal{H}_0$ that the time series of residuals for each asset are independent:

$$\mathcal{H}_0 : \epsilon_{\cdot,1}, \epsilon_{\cdot,2}, \ldots, \epsilon_{\cdot,p} \in \mathbb{R}^T \text{ are jointly independent.} \tag{2.2}$$

Sections 2.1 and 2.2 now give some intuition explaining why naive permutation and bootstrap methods for testing $\mathcal{H}_0$ can yield false positive rates of up to 100%, as shown by Figure 3. The main ideas in Section 2.1 and 2.2, respectively, are that (i) the estimated OLS residuals $\hat{\epsilon}^{\text{ols}}$ from Eq. 1.3 not satisfy the same independence properties as the true residuals, making permutation-based inference challenging, and (ii) our problem setting is too high-dimensional for the bootstrap to perform well (Bickel and Freedman, 1983; Karoui and Purdom, 2018).

2.1 Naive residual permutation tests are invalid

For simplicity, we assume for this section that the residuals $\epsilon_{1,j}, \ldots, \epsilon_{T,j} \overset{\text{i.i.d.}}{\sim} P_j$ for each asset are drawn i.i.d. from an asset-specific distribution. Under $\mathcal{H}_0$, the residuals for different assets are independent.

Thus, the i.i.d. assumption plus $\mathcal{H}_0$ together imply that separately permuting the residuals of each asset does not change the joint law of all of the residuals:

$$\epsilon := \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,p} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,p} \\ \epsilon_{3,1} & \epsilon_{3,2} & \dots & \epsilon_{3,p} \\ \vdots & \vdots & \dots & \vdots \\ \epsilon_{T,1} & \epsilon_{T,2} & \dots & \epsilon_{T,p} \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} \epsilon_{\pi_1(1),1} & \epsilon_{\pi_2(1),2} & \dots & \epsilon_{\pi_p(1),p} \\ \epsilon_{\pi_1(2),1} & \epsilon_{\pi_2(2),2} & \dots & \epsilon_{\pi_p(2),p} \\ \epsilon_{\pi_1(3),1} & \epsilon_{\pi_2(3),2} & \dots & \epsilon_{\pi_p(3),p} \\ \vdots & \vdots & \dots & \vdots \\ \epsilon_{\pi_1(T),1} & \epsilon_{\pi_2(T),2} & \dots & \epsilon_{\pi_p(T),p} \end{bmatrix}, \tag{2.3}$$

where above, $\pi_1, \dots, \pi_p : [T] \to [T]$ are arbitrary permutations applied to the p columns of ϵ. Thus, if we observed ϵ, we could easily design a permutation test of $\mathcal{H}_0$ as follows.

1. Permute each of the columns of ϵ uniformly at random, and repeat this R times, yielding permuted matrices $\epsilon^{(1)}, \dots, \epsilon^{(R)} \in \mathbb{R}^{T \times p}$.
2. Let $S(\epsilon)$ be any test statistic, such as the maximum empirical correlation among the residuals. Compute a p-value by comparing the value of $S(\epsilon)$ to $S(\epsilon^{(1)}), \dots, S(\epsilon^{(R)})$:

$$p_{\mathrm{val}} := \frac{1 + \sum_{r=1}^{R} \mathbb{I}(S(\epsilon) \le S(\epsilon^{(r)}))}{R+1}, \tag{2.4}$$

   where Equation 2.3 guarantees that this is a finite-sample p-value testing $\mathcal{H}_0$.

Although we do not observe the residuals ϵ, a "naive residual permutation test" would simply plug in the OLS estimate $\hat{\epsilon}^{\mathrm{ols}} \in \mathbb{R}^{T \times p}$ in place of ϵ, where $\hat{\epsilon}^{\mathrm{ols}}_t = H^{\mathrm{ols}}_t \epsilon_t \in \mathbb{R}^p$ for the projection matrix H^{ols}_t in Eq. 1.3. Unfortunately, this strategy will not work: while the coordinates of ϵ_t are independent under $\mathcal{H}_0$, the coordinates of $\hat{\epsilon}^{\mathrm{ols}}$ are certainly not. Indeed, the covariance matrix $\mathrm{Cov}(\hat{\epsilon}^{\mathrm{ols}}_t) = H^{\mathrm{ols}}_t \mathrm{Cov}(\epsilon_t) H^{\mathrm{ols}}_t$ has a reduced rank of at most $p - k$. As a result, naively replacing ϵ with $\hat{\epsilon}^{\mathrm{ols}}$ will violate Eq. 2.3—in particular, the columns of $\hat{\epsilon}^{\mathrm{ols}}$ will look much more correlated than the permuted version of $\hat{\epsilon}^{\mathrm{ols}}$, even under the null. Indeed, Figure 3(a) uses the real exposures from the BFRE model to show that this "naive permutation test" may cause an unacceptably high false positive rate.

2.2 Naive bootstrap methods are invalid

Another naive way to test $\mathcal{H}_0$ would be to use the nonparametric bootstrap to compute a Z-statistic based on $S(\hat{\epsilon}^{\mathrm{ols}})$. This strategy does not adjust the estimated residuals to force them to satisfy the null—rather, it reframes the hypothesis testing problem as an estimation problem. In particular, suppose that $S(\hat{\epsilon}^{\mathrm{ols}})$ is an estimate of some parameter θ which equals zero under $\mathcal{H}_0$. For example, in Section 1.2, S^{ols}_t is a (biased) estimate of the true mean absolute maximum correlation of the residuals, which is zero under $\mathcal{H}_0$. We hope to use the bootstrap to debias and standardize $S(\hat{\epsilon})$, allowing us to test whether $\theta = 0$, thus testing the null. There are many ways to apply the bootstrap, but perhaps the simplest is as follows:

1. Resample T rows from $\hat{\epsilon}^{\mathrm{ols}}$ uniformly at random and with replacement.[1]
2. Repeat this B times, yielding B bootstrapped residual matrices $\hat{\epsilon}^{\mathrm{ols},(1)}, \dots, \hat{\epsilon}^{\mathrm{ols},(B)} \in \mathbb{R}^{T \times p}$.
3. Compute a bootstrap bias estimate for $S(\hat{\epsilon}^{\mathrm{ols}})$ as well as a Z-statistic which is intended to have zero mean and unit variance under the null:

$$\widehat{\mathrm{Bias}} = \frac{1}{B} \sum_{b=1}^{B} S(\hat{\epsilon}^{\mathrm{ols},(b)}) - \theta_{\mathrm{BS}} \quad \text{and} \quad Z_{\mathrm{BS}} = \frac{S(\hat{\epsilon}^{\mathrm{ols}}) - \widehat{\mathrm{Bias}}}{\sqrt{\widehat{\mathrm{Var}}(\{S(\hat{\epsilon}^{\mathrm{ols},(b)})\}_{b=1}^{B})}}, \tag{2.5}$$

[1] Since the test statistic only depends on $\mathbf{Y}$ through $\hat{\epsilon}^{\mathrm{ols}}$, the residual bootstrap is in this case identical to the pairs bootstrap, which resamples pairs of exposures and returns $\{(L_t, Y_t)\}_{t=1}^{T}$.

where θ_{BS} is the value of the parameter θ calculated for the bootstrap empirical distribution. (See Appendix C.3 for a discussion of how to compute θ_{BS} in our setting.)

Unfortunately, inference based on the procedure above can be highly misleading. The main reason is that the estimated residuals $\hat{\epsilon}_t^{\text{ols}} \in \mathbb{R}^p$ are "high-dimensional" vectors in the sense that the number of assets p is usually not negligible compared to the number of observations T. As a result, the bootstrap distribution of the estimated residuals $\hat{\epsilon}_t^{\text{ols}}$ may not accurately approximate the law of the residuals ϵ_t (or even the true law of the estimated residuals $\hat{\epsilon}_t^{\text{ols}}$) (Bickel and Freedman, 1983; Karoui and Purdom, 2018). Thus, the bias and variance estimates can be highly inaccurate, since they are based on a bootstrap distribution which differs substantially from the true data-generating process. And despite recent work on high-dimensional bootstraps (see Chernozhukov et al., 2023, for review), we are not aware of existing bootstrap methods with inferential guarantees for our problem.

Empirically, in the semisynthetic simulations in Figure 3(b), the bootstrap bias estimate (≈ 0.06) is over three times smaller than the true bias of the test statistic (≈ 0.2). As a result, Figure 3(a) shows that the bootstrap Z-statistics are highly inaccurate and have an average value of 25 (while we would expect to see an average of ≈ 0 if the test were performing well), leading to essentially a 100% false positive rate.

3 Methodology

3.1 Main idea

As discussed in Section 2.1, the key challenge in developing a permutation test for $\mathcal{H}_0$ is that the OLS residual estimates $\hat{\epsilon}^{\text{ols}}$ do not satisfy the same independence properties as the true residuals ϵ. The main idea behind the mosaic permutation test is to introduce a new estimator $\hat{\epsilon}$ which exactly preserves some of the independence properties of ϵ. To ease readability and build intuition, this subsection introduces the simplest possible variant of the mosaic permutation test. Section 3.2 then introduces the mosaic permutation test in full generality. However, Section 3.2 is self-contained, so readers may skip to Section 3.2 if they wish.

For simplicity of exposition, we temporarily make two simplifying assumptions for this subsection only:

- The residuals $\epsilon_{1,j}, \dots, \epsilon_{T,j} \overset{\text{i.i.d.}}{\sim} P_j$ for each asset are drawn i.i.d. from an asset-specific distribution.
- The exposures $L_t = L \in \mathbb{R}^{p \times k}$ do not change with time.

The main idea is to split the assets into two groups, $G_1, G_2 \subset [p]$, and estimate the residuals separately for each group. This ensures that under the null, the *estimated* residuals for the assets in G_1 and G_2 are independent, so we can separately permute the residuals for those assets. To be precise, consider the procedure below:

1. Partition the set of assets into two groups $[p] = G_1 \cup G_2$ with $G_1 \cap G_2 = \emptyset$. For now, take $G_1 = \{1, \dots, \lfloor p/2 \rfloor\}$ and $G_2 = \{\lfloor p/2 \rfloor + 1, \dots, p\}$.
2. For $i \in \{1, 2\}$, let $\hat{\epsilon}_{t,G_i}$ denote the OLS estimate of $\epsilon_{t,G_i} \in \mathbb{R}^{|G_i|}$ based on $Y_{t,G_i} \in \mathbb{R}^{|G_i|}$, the returns of the assets in group G_i at time t. Formally, let $H_i = (I_{|G_i|} - L_{G_i}(L_{G_i}^\top L_{G_i})^{-1} L_{G_i}^\top)$ be the OLS projection matrix based on L_{G_i}, the exposures for the assets in G_i. Then we have that

$$\hat{\epsilon}_{t,G_i} \coloneqq H_i Y_{t,G_i} = H_i \epsilon_{t,G_i}, \tag{3.1}$$

   where the equality above holds because $H_i L_{G_i} X_{t,G_i} = 0$ by construction of H_i.

Remark 1. In each regression above, the parameters are the factor returns $X_t \in \mathbb{R}^k$ and the "number of observations" is the number of stocks $|G_i|$ in group $i \in \{1, 2\}$. By splitting the assets into two groups, we reduce the number of "observations" in each regression by a factor of two, leading to higher estimation error of the residuals ϵ. However, in most typical applications, the total number of stocks p is much larger than the number of factors k, so we will still obtain reasonably good estimates $\hat{\epsilon}$. That said, in rare settings where $p < 2k$, then for at least one $i \in \{1, 2\}$ we have that $k > |G_i|$, in which case $\hat{\epsilon}_{t,G_i} = 0$ deterministically and this procedure will be powerless.

3. Let $\hat{\epsilon}$ denote the appropriate concatenation of $\{(\hat{\epsilon}_{t,G_1}, \hat{\epsilon}_{t,G_2})\}_{t=1}^T$:

$$\hat{\epsilon} := \begin{bmatrix} \hat{\epsilon}_{1,G_1} & \hat{\epsilon}_{1,G_2} \\ \hat{\epsilon}_{2,G_1} & \hat{\epsilon}_{2,G_2} \\ \vdots & \vdots \\ \hat{\epsilon}_{T,G_1} & \hat{\epsilon}_{T,G_2} \end{bmatrix} \in \mathbb{R}^{T \times p}. \tag{3.2}$$

Since the groups of stocks G_1, G_2 are disjoint, under $\mathcal{H}_0$ we have that $\hat{\epsilon}_{t,G_1} = H_1 \epsilon_{t,G_1}$ and $\hat{\epsilon}_{t,G_2} = H_2 \epsilon_{t,G_2}$ are independent. Furthermore, $\{\epsilon_{t,G_i}\}_{t=1}^T$ are i.i.d. for each $i \in \{1, 2\}$. Therefore, we can separately permute $\{\epsilon_{t,G_1}\}_{t=1}^T$ and $\{\epsilon_{t,G_2}\}_{t=1}^T$ without changing the law of $\hat{\epsilon}$, as illustrated below:

$$\hat{\epsilon} := \begin{bmatrix} \hat{\epsilon}_{1,G_1} & \hat{\epsilon}_{1,G_2} \\ \hat{\epsilon}_{2,G_1} & \hat{\epsilon}_{2,G_2} \\ \vdots & \vdots \\ \hat{\epsilon}_{T,G_1} & \hat{\epsilon}_{T,G_2} \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} \hat{\epsilon}_{\pi_1(1),G_1} & \hat{\epsilon}_{\pi_2(1),G_2} \\ \hat{\epsilon}_{\pi_1(2),G_1} & \hat{\epsilon}_{\pi_2(2),G_2} \\ \vdots & \vdots \\ \hat{\epsilon}_{\pi_1(T),G_1} & \hat{\epsilon}_{\pi_2(T),G_2} \end{bmatrix} := \hat{\epsilon}^{(\pi)} \tag{3.3}$$

where $\pi_1, \pi_2 : [T] \to [T]$ are any permutations. Due to the complex notation, the idea is best understood visually. A color-assisted illustration of this equality is given below with $T = 6$ observations.

$\hat{\epsilon} =$

p assets split into two subsets	
$\hat{\epsilon}_{1,G_1}$	$\hat{\epsilon}_{1,G_2}$
$\hat{\epsilon}_{2,G_1}$	$\hat{\epsilon}_{2,G_2}$
$\hat{\epsilon}_{3,G_1}$	$\hat{\epsilon}_{3,G_2}$
$\hat{\epsilon}_{4,G_1}$	$\hat{\epsilon}_{4,G_2}$
$\hat{\epsilon}_{5,G_1}$	$\hat{\epsilon}_{5,G_2}$
$\hat{\epsilon}_{6,G_1}$	$\hat{\epsilon}_{6,G_2}$

$\overset{\mathrm{d}}{=}$

$\hat{\epsilon}_{\cdot,G_1}, \hat{\epsilon}_{\cdot,G_2}$ are separately permuted	
$\hat{\epsilon}_{5,G_1}$	$\hat{\epsilon}_{3,G_2}$
$\hat{\epsilon}_{1,G_1}$	$\hat{\epsilon}_{6,G_2}$
$\hat{\epsilon}_{4,G_1}$	$\hat{\epsilon}_{1,G_2}$
$\hat{\epsilon}_{3,G_1}$	$\hat{\epsilon}_{2,G_2}$
$\hat{\epsilon}_{6,G_1}$	$\hat{\epsilon}_{5,G_2}$
$\hat{\epsilon}_{2,G_1}$	$\hat{\epsilon}_{4,G_2}$

(3.4)

Above, the two groups are shown in different colors, and the shading of each cell denotes its original position in time—for this reason, the shadings in the right panel indicate that $\hat{\epsilon}_{\cdot,G_1}$ and $\hat{\epsilon}_{\cdot,G_2}$ have been separately permuted. Throughout the paper, we will use these figures as much as we can to make the mathematical notation easier to understand.

After sampling M permutations uniformly at random to create M new estimated residual matrices $\tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}$, we can compute a valid p-value using any test statistic $S : \mathbb{R}^{T \times p} \to \mathbb{R}$:

$$p_{\text{val}} := \frac{1 + \sum_{r=1}^R \mathbb{I}(S(\hat{\epsilon}) \le S(\tilde{\epsilon}^{(r)}))}{R + 1}. \tag{3.5}$$

For example, $S(\hat{\epsilon})$ could measure the maximum absolute correlation of a (regularized) estimate of the covariance matrix of ϵ_t. We discuss the choice of test statistic in more detail in Sections 4 and 5. The key result is that this p-value is valid for any test statistic.

Theorem 3.1. *p_{val} in Eq. 3.5 is an exact p-value testing $\mathcal{H}_0$ assuming $\epsilon_1, \ldots, \epsilon_T \overset{\text{i.i.d.}}{\sim} P_\epsilon$ are i.i.d.*

This simple procedure already has many desirable properties, but unfortunately, it lacks stability due to the choice of G_1 and G_2. For example, suppose that in truth, the idiosyncratic returns of stocks 1 and 2 are highly correlated. We will have no power to detect this if both stocks are placed in the same group because the estimated idiosyncratic returns $\hat{\epsilon}_{\cdot,1}, \hat{\epsilon}_{\cdot,2}$ for the first two stocks will never be "separated" by different permutations. We address this problem in the next section.

3.2 The mosaic permutation test

We now introduce the general mosaic permutation test, which is more powerful and stable than the simple method in Section 3.1. As an added benefit, we will also make the test more robust to autocorrelation and distribution drift among the residuals $\epsilon_1, \ldots, \epsilon_T$.

In Section 3.1, we separated the asset returns $\mathbf{Y} \in \mathbb{R}^{T \times p}$ into two disjoint groups, computed residual estimates $\hat{\epsilon}$ separately for each group, and then permuted within each group. Now, we partition the data $\mathbf{Y}$ into an arbitrary number M of rectangles along both axes. To make this precise, for $m = 1, \ldots, M$, let $B_m \subset [T]$ denote a subset or "batch" of observations and $G_m \subset [p]$ denote a subset or "group" of assets. We say $\{(B_m, G_m)\}_{m=1}^M$ is a *tiling* if for every timepoint t and asset j, there is exactly one pair (B_m, G_m) such that $t \in B_m$ and $j \in G_m$. See Figure 5 for an illustration of this definition.

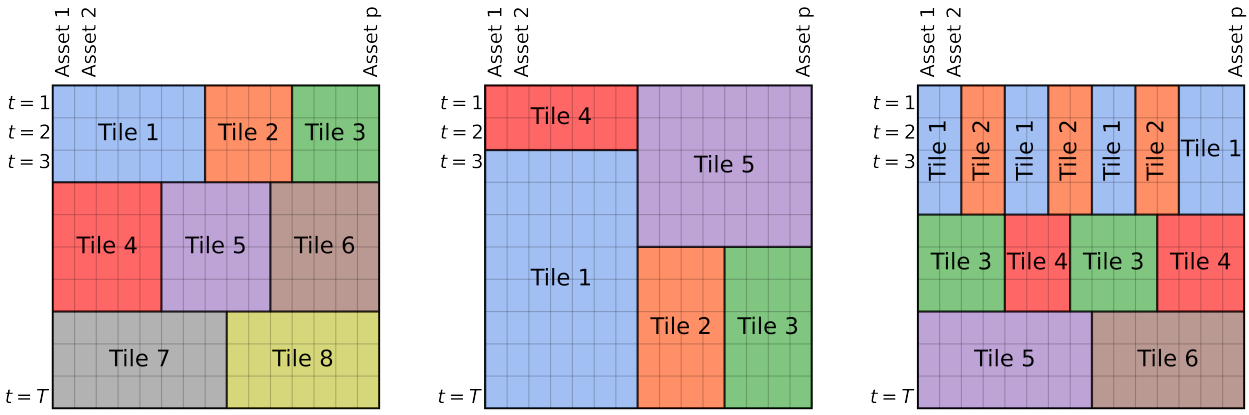


Figure 5: This figure shows three examples of tilings of the data matrix $\mathbf{Y}$. The right-most example emphasizes that each tile need not be contiguous in the initial ordering of the features.

We will soon discuss how to choose the tiling. For now, given an arbitrary tiling, we refer to the submatrices $\mathbf{Y}_{(m)} = \mathbf{Y}_{B_m, G_m}$ and $\epsilon_{(m)} := \epsilon_{B_m, G_m}$ as the mth *tiles* of the matrices $\mathbf{Y}$ and ϵ, respectively. Before presenting the mosaic permutation test, we make two assumptions, generalizing the assumptions from Section 3.1.

First, in Section 3.1, we assumed that the exposures L_t did not change with time. Now, we ask that the analyst chooses the tiles such that the exposure matrices $\{L_t\}_{t=1}^T$ may change across tiles, but not within tiles. For example, in our application to the BFRE model, the exposures change every week. As a result, we choose the tiles such that each tile contains data from only one week. If L_t takes unique values at every observation, then testing $\mathcal{H}_0$ is possible but requires a slightly more sophisticated technique introduced in Section 5.3.

Assumption 3.1 (Constant L_t within tiles)**.** *The exposures $\{L_t\}_{t \in B_m}$ in each tile are all equal.*

Second, in Section 3.1, we assumed that the idiosyncratic returns $\epsilon_1, \dots, \epsilon_T$ were i.i.d. We now relax this assumption to allow for a large degree of nonstationarity among the residuals. In particular, we assume that each asset's returns are exchangeable *within* tiles, but not necessarily between tiles.

Assumption 3.2 (Local exchangeability)**.** *For every asset $j \in [p]$, we assume the following. Let $\pi : [T] \to [T]$ be any permutation such that $\mathbf{Y}_{t,j}$ and $\mathbf{Y}_{\pi(t),j}$ are always in the same tile. Then*

$$(\epsilon_{1,j}, \dots, \epsilon_{T,j}) \stackrel{\mathrm{d}}{=} (\epsilon_{\pi(1),j}, \dots, \epsilon_{\pi(T),j}). \tag{3.6}$$

Assumption 3.2 allows the distribution of the residuals to drift between tiles, making this assumption much weaker than the i.i.d. assumption in Theorem 3.1. Indeed, this assumption is related to the motivation for many classical procedures for time series data, such as the block bootstrap (Kunsch, 1989). Armed with these assumptions, Algorithm 1 defines the mosaic permutation test.

Algorithm 1: The mosaic permutation test.

Inputs: Asset returns $\mathbf{Y} \in \mathbb{R}^{T \times p}$, exposures $L_t \in \mathbb{R}^{p \times k}$ for $t \in [T]$, tiles $\{(B_m, G_m)\}_{m=1}^M$ and a test statistic $S : \mathbb{R}^{T \times p} \to \mathbb{R}$.

Step 1: For each tile $m = 1, \dots, M$, we let $\hat{\epsilon}_{(m)}$ denote the OLS estimate of $\epsilon_{(m)}$ using only the data in $\mathbf{Y}_{(m)}$. Precisely, let $L_{(m)} \in \mathbb{R}^{|G_m| \times k}$ denote the exposures for the assets in the mth tile (note by Assumption 3.1 that the exposures are constant over time within the tile). Let H_m be the OLS projection matrix based on $L_{(m)}$:

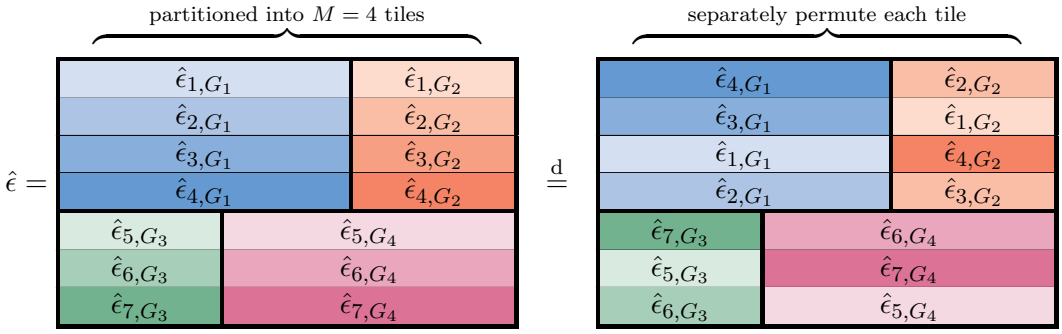
$$H_m \coloneqq (I_{|G_m|} - L_{(m)}(L_{(m)}^\top L_{(m)})^{-1} L_{(m)}^\top).$$

Then we define

$$\hat{\epsilon}_{(m)} \coloneqq \mathbf{Y}_{(m)} H_m = \epsilon_{(m)} H_m \tag{3.7}$$

and $\hat{\epsilon} \in \mathbb{R}^{T \times p}$ denotes the appropriate concatenation of the tiles $\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(M)}$. In other words, $\hat{\epsilon}$ is defined such that $\hat{\epsilon}_{B_m, G_m} \coloneqq \hat{\epsilon}_{(m)}$ for $m \in [M]$.

Step 2: For $r = 1, \dots, R$, randomly reorder the rows within each tile and let $\tilde{\epsilon}^{(r)} \in \mathbb{R}^{T \times p}$ denote the resulting matrix. Visually, this is depicted below with $T = 7$ observations and $M = 4$ tiles:



(3.8)

To define this mathematically, let $P_m^{(r)} \in \mathbb{R}^{|B_m| \times |B_m|}$ denote a uniformly random permutation matrix for $m \in [M], r \in [R]$. Then $\tilde{\epsilon}^{(r)}$ is defined such that $\tilde{\epsilon}^{(r)}_{B_m, G_m} \coloneqq P_m^{(r)} \hat{\epsilon}_{(m)}$ for $m \in [M]$.

Step 3: For any test statistic $S : \mathbb{R}^{T \times p} \to \mathbb{R}$, compute the p-value

$$p_{\mathrm{val}} \coloneqq \frac{1 + \sum_{r=1}^R \mathbb{I}(S(\hat{\epsilon}) \le S(\tilde{\epsilon}^{(r)}))}{R + 1}. \tag{3.9}$$

Remark 2. If the mth tile contains data from $t_m = |B_m|$ timepoints, then Eq. 3.7 is equivalent to running t_m separate cross-sectional regressions to compute each row of $\hat{\epsilon}_{(m)}$. Precisely,

$$\hat{\epsilon}_{t,G_m} \coloneqq H_m Y_{t,G_m} = H_m \epsilon_{t,G_m}.$$

In other words, the "number of observations" in this regression is the number of stocks in G_m, and the parameters are the factor returns $X_t \in \mathbb{R}^k$.

Figure 2 gives a simple illustration of the mosaic permutation test with $M = 10$ tiles; note that the initial method introduced in Section 3.1 is an example of this procedure with $M = 2$ tiles. Theorem 3.2 states that p_{val} is a valid p-value assuming only Assumptions 3.1-3.2. We emphasize that these assumptions allow for (i) the residuals ϵ and factors $\{X_t\}_{t=1}^T$ to be arbitrarily heavy-tailed, (ii) the factors $\{X_t\}_{t=1}^T$ to be arbitrarily nonstationary and autocorrelated, and (iii) the residuals $\{\epsilon_t\}_{t=1}^T$ to be nonstationary between batches, and (iv) the use of any test statistic $S(\cdot)$.

Theorem 3.2. *Suppose Assumptions 3.1-3.2 hold. Then under $\mathcal{H}_0$, Eq. 3.9 defines a valid p-value satisfying $\mathbb{P}(p_{\text{val}} \leq \alpha) \leq \alpha$ for any $\alpha \in (0, 1)$.*

A formal proof of Theorem 3.2 is given in Appendix A. However, the main idea is that the residuals for the tiles $\{\hat{\epsilon}_{(m)}\}_{m=1}^M$ are estimated using separate data. Thus, under the null $\mathcal{H}_0$ that the columns of ϵ are independent, the tiles have exchangeable rows. As a result, separately reordering the rows within each tile does not change the joint law of the full estimated residual matrix $\hat{\epsilon}$. Formally, if $P_1 \in \{0,1\}^{|B_1|\times|B_1|}, \ldots, P_M \in \{0,1\}^{|B_M|\times|B_M|}$ are a sequence of permutation matrices, we have that

$$(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(M)}) \overset{\text{d}}{=} (P_1\hat{\epsilon}_{(1)}, \ldots, P_M\hat{\epsilon}_{(M)}). \tag{3.10}$$

Remark 3 (Regularization)**.** Theorem 3.2 allows the use of arbitrary test statistics $S(\hat{\epsilon})$, including regularized estimates of $\text{Cov}(\epsilon)$. However, we require that the analyst use unregularized OLS regressions in each tile to estimate $\hat{\epsilon}$. The reason for this is that regularized (e.g.) ridge estimates of the residuals ϵ will not project out the influence of the factor returns $\{X_t\}_{t\in[T]}$, causing violations of Eq. 3.10 and potentially leading to inflated false positives. For example, in Eq. 3.8, if one used regularized regressions instead of OLS regressions to compute $\hat{\epsilon}$, then $\hat{\epsilon}_{1,G_1}$ and $\hat{\epsilon}_{1,G_2}$ would not be fully independent since they would both depend to some extent on X_1. That said, incorporating regularized estimates of the residuals is a promising direction for future work (see Section 7).

3.3 A default choice of tiling

We recommend choosing the tiling by doing the following:

1. First, partition the time points into $[T] = B_1 \cup \cdots \cup B_I$ into $I \in \mathbb{N}$ batches. By default, we take each batch to be 10 consecutive time points, so $B_1 = \{1, \ldots, 10\}, B_2 = \{11, \ldots, 20\}$, etc. That said, if necessary, one may make the batches smaller to guarantee that the exposures are constant within each batch (satisfying Assumption 3.1).
2. For each batch $i = 1, \ldots, I$, randomly partition the assets into D groups $[p] = G_{i,1} \cup \cdots \cup G_{i,D}$ of (roughly) equal size. We recommend setting $D = \max\left(2, \left\lceil \frac{p}{5k} \right\rceil\right)$.
3. We let $\{(B_i, G_{i,k}) : d \in [D], i \in [I]\}$ be the final set of $M = I \cdot D$ tiles. For example, in Figure 5, the tilings in the left and right panels are of this form (while the one in the middle panel is not).

Above, we use small batch sizes of ≈ 10 observations because smaller contiguous batches are more robust to nonstationarity and autocorrelation among the residuals. Furthermore, using more batches increases the stability of the test and decreases the likelihood that any one random partition of the assets dramatically affects the results.

Note also that the choice of D balances the following trade-off. On the one hand, using larger D increases the probability that any two assets are in different tiles and thus that their estimated idiosyncratic returns can be "separated" by permutations, since there is a $\approx \frac{1}{D}$ chance that any two assets lie in the same group. On the other hand, as per Remark 2, we must separately estimate the value of the factors $X_t \in \mathbb{R}^k$ within each tile using a linear regression with $\frac{p}{D}$ observations. Increasing D will reduce the number of observations per regression and increase the estimation error of $\hat{\epsilon}$. Thus, choosing $D = \max(2, \lfloor \frac{p}{5k} \rfloor)$ maximizes the number of tiles subject to the constraint that there are 5 times as many observations as there are parameters in each regression used to estimate $\hat{\epsilon}$.

4 Application to the BFRE model

We now apply the mosaic permutation test to the Blackrock Fundamental Equity Risk (BFRE) model as described in Section 1.2. We ask whether the BFRE model accurately describes correlations among asset returns in three sectors of the US economy: energy, financials, and healthcare. We also analyze six additional sectors, but for brevity, we present these additional results in Appendix B.2. However, the results below are representative of our findings in the appendix.

Overall, we find that the BFRE model explains the majority of correlations among assets. However, we find evidence of a missing persistent factor among real estate stocks and a missing transient factor among healthcare stocks post-COVID. We do not find consistent evidence against the null in the energy sector. These findings are supported by four analyses, detailed in the next four sections:

- Our main analysis straightforwardly applies the mosaic permutation test to each sector.
- Our second analysis confirms in an ablation test that removing a fraction of the existing factors from the BFRE model leads to strong evidence against the null.
- Our third analysis analyzes the degree to which we can consistently improve the BFRE model, as measured by a mosaic estimate of out-of-sample predictive performance.
- Our final analysis shows that the unexplained correlations among financial assets are largely concentrated among real estate assets.

4.1 Main analysis

We now present our main analysis of the BFRE model. Methodologically, we use the default choice of tiling discussed in Section 3.3, and we use the mean maximum absolute correlation (MMC) test statistic from Section 1.1. We use this test statistic because it is simple and interpretable, although we shall soon discuss other choices.

Figure 4 shows the value of the test statistic and null quantile for these three sectors, and Figure 6 plots the mosaic Z-statistic $Z = \Phi^{-1}(1 - p_{\text{val}})_+$, where Φ^{-1} denotes the inverse CDF of a standard normal distribution. Z is distributed as the positive part of a standard normal under the null, so large values of Z are evidence against the null.

The findings are qualitatively different for different sectors. For instance, in the energy sector, we generally do not find statistically significant evidence against the null, whereas in the healthcare sector, we primarily find evidence against the null after the COVID-19 pandemic began, suggesting that we

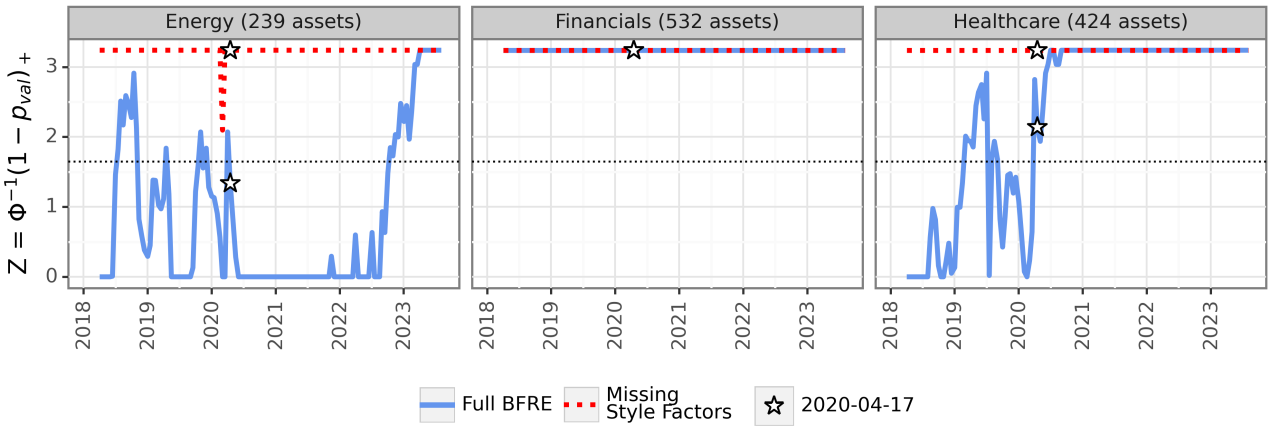


Figure 6: This figure plots the mosaic z-statistics $\Phi^{-1}(1 - p_{\text{val}})_+$ computed for the analysis in Figure 4 over 200 time points. It also shows the Z-statistics for the ablation analysis in Figure 7. In all instances, we apply a Bonferroni correction for the $m = 3$ sectors, but we do not apply a multiplicity correction across all 200 time points; thus, these z-statistics are only marginally valid. The dotted black line denotes the marginal significance threshold (1.64) for $\alpha = 0.05$. Note that the maximum z-statistic value is ≈ 3.24, corresponding to the minimum adjusted p-value of $\frac{3}{5001}$, since the p-values were computed using $R = 5000$ randomizations.

may be detecting a transient factor. In the financial sector, we find highly statistically significant evidence of violations of the null $\mathcal{H}_0$ at all time points, and the resulting p-values are highly significant ($p < 0.001$), suggesting that we are detecting a persistent factor. That said, it is not clear if this represents a large effect size. Indeed, heuristically, the observed value of the test statistic S_t is usually quite close to its null threshold $Q_{0.95}(S_t)$—in healthcare and financials, respectively, the observed MMC statistics are at most 3% and 7% larger than their estimated null quantiles. To investigate this further, we perform three additional analyses, detailed in the next three sections.

4.2 Ablation test

We now perform an ablation test which shows that after removing a subset of factors from the BFRE model, (a) the mosaic p-values become more significant and (b) the difference between S_t and its null quantile substantially increases. This suggests (heuristically) that the BFRE model accounts for the most significant correlations among assets, even though it does not fit perfectly.

To precisely describe this analysis, note that the BFRE model contains two types of factors. First, it contains *industry* factors, where asset j's exposure to (e.g.) the banking industry factor is simply the binary indicator of whether asset j is a bank. Second, it contains twelve *style* factors, such as "size," where a corporation's exposure to the "size" factor is simply a measure of the size of the corporation. Figure 7 replicates our previous analysis of the BFRE model after removing the style factors, and it shows that the difference between the test statistic and its estimated null quantile generally doubles compared to the analysis of the full BFRE model. Figure 6 shows that the mosaic p-values in this analysis are nearly all equal to their minimum possible value at all time points, giving highly significant evidence against the null.

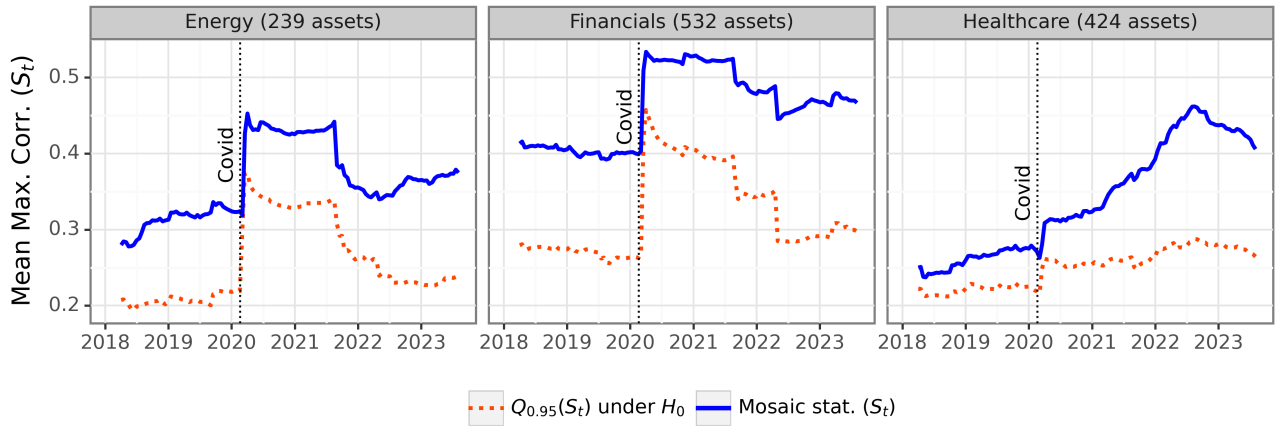


Figure 7: This figure replicates the analysis in Figure 1, except now we perform an ablation test where we remove the twelve style factors from the BFRE model.

4.3 Improving the model

The main contribution of our paper is a methodology for testing the goodness-of-fit of a factor model. That said, we recognize that in many applications, another important question is whether one can *improve* a model. We now illustrate how the mosaic permutation test can help answer this question.

Consider a simple alternative where the BFRE model is missing at least one component:

$$Y_t = L_t X_t + v Z_t + \epsilon_t, \tag{4.1}$$

where $L_t \in \mathbb{R}^{p \times k}$ are the pre-existing factor exposures in the BFRE model, X_t are the BFRE factors, $Z_t \in \mathbb{R}$ is an additional factor and $v \in \mathbb{R}^p$ is an unknown (missing) factor. We ask: can we estimate v such that the model fit is improved? Or does including a new set of exposures v degrade the model performance by adding too much additional variance?

To measure model performance, we split the data into two folds. On the first fold, we estimate new exposures $\hat{v}$. On the second fold, we check whether $\hat{v}$ allows us to better predict each residual $\hat{\epsilon}_{t,j}$ from the residuals in the other tiles. (Note that under the null, $\hat{\epsilon}_{t,j}$ is independent of $\hat{\epsilon}_{t,j'}$ for all assets $j' \in [p]$ which are in a different tile from j at time t, so this prediction task is hopeless unless the null is violated.) This test statistic is inspired by, although not identical to, the *bi-cross validation* statistic from Owen and Wang (2016); Owen and Perry (2009); Perry (2009). Finally, we check statistical significance using the mosaic permutation test. The details of this analysis are described below.

Step 1: estimating v. For each sector, we construct several estimators $\hat{v}_0, \ldots, \hat{v}_I$ of v using the first $T_0 \approx 1400$ observations, from 2014 through October 2020. Precisely, let $\hat{C} \in \mathbb{R}^{p \times p}$ denote the empirical correlation matrix of the mosaic residual estimates $\{\hat{\epsilon}_t\}_{t=1}^{T_0}$. Our first estimate $\hat{v}_0$ is simply the top eigenvector of $\hat{C}$. However, in many settings, v will be sparse, and incorporating assumptions about sparsity could lead to higher power. Since we do not know the sparsity level of v a priori, we approximately solve the sparse PCA objective for various sparsity levels ℓ:

$$\hat{v} \approx \max_{\|v\|_2 = 1} v^T \hat{C} v \text{ s.t. } \|v\|_0 \leq \ell, \tag{4.2}$$

where above, $\|v\|_0 = |\{i \in [p] : v_i \neq 0\}|$ counts the number of nonzero entries of v. Exactly solving this optimization problem is computationally intractable, but we use a simply greedy approximation

detailed in Appendix C.2. For each sector, we compute estimators $\hat{v}_1, \ldots, \hat{v}_I$ for $I = 10$ values of ℓ, evenly spaced between 20 and p.

Step 2: measuring out-of-sample performance. For each $\hat{v}_i$, we compute a mosaic estimate of the out-of-sample error of the new model. Precisely, fix any time point $t > T_0$ and asset j in the mth tile. We check whether $\hat{v}_i$ allows us to predict the mosaic residual $\hat{\epsilon}_{t,j}$ from $\hat{\epsilon}_{t,-G_m}$, the set of residuals at time t which are not in the same tile as $\hat{\epsilon}_{t,j}$. To do this, we first compute an OLS estimate of Z_t, the missing factor return, using $\hat{\epsilon}_{t,-G_m}$:

$$\hat{Z}_t^{(i,j)} = \frac{\hat{\epsilon}_{t,-G_m}^T \hat{v}_{i,-G_m}}{\|\hat{v}_{i,-G_m}\|_2^2} \in \mathbb{R}.$$

Then, we set $\hat{\gamma}_{t,j}^{(i)}$ to be the out-of-sample OLS estimate of $\hat{\epsilon}_{t,j}$ based on $\hat{Z}_t^{(i,j)}$:

$$\hat{\gamma}_{t,j}^{(i)} = \hat{v}_{i,j} \cdot \hat{Z}_t^{(i,j)}.$$

Note that $\hat{\gamma}_{t,j}^{(i)}$ depends only on $\hat{\epsilon}_{t,-G_m}$ and the first fold of the dataset, so if the two folds of the dataset are independent and $\mathcal{H}_0$ holds, then $\hat{\gamma}_{t,j}^{(i)}$ is independent of $\hat{\epsilon}_{t,j}$. In contrast, if $\mathcal{H}_0$ is violated and $\hat{v}_i$ explains additional correlations among the assets, then $\hat{\gamma}_{t,j}^{(i)}$ should predict $\hat{\epsilon}_{t,j}$. As an aggregate measure of model performance, we compute the out-of-sample R^2 of these new predictions:

$$\hat{r}_i^2 = 1 - \frac{\sum_{t=T_0+1}^{T} \sum_{j=1}^{p} \left(\hat{\gamma}_{t,j}^{(i)} - \hat{\epsilon}_{t,j}\right)^2}{\sum_{t=T_0+1}^{T} \sum_{j=1}^{p} \hat{\epsilon}_{t,j}^2}.$$

Following Owen and Wang (2016), we refer to this as a mosaic *bi-cross validation* (BCV) R^2 value. Our final test statistic is the maximum of the mosaic BCV estimates $\hat{r}_i^2$ over all of the sparse-PCA estimates $\hat{v}_0, \ldots, \hat{v}_I$:

$$S = \max(\hat{r}_0^2, \ldots, \hat{r}_I^2). \tag{4.3}$$

By taking the maximum among the out-of-sample R^2 values, we hope to gain power no matter the underlying sparsity level of any missing factor exposures v. We compute this statistic in sliding windows of 350 days on the second fold of data, which ranges from October 2020 through October 2023, and we check for statistical significance using the mosaic permutation test.

Figures 8 show the results. As expected, the maximum BCV R^2 value is negative and statistically insignificant for the energy sector, and it is positive (usually $\approx 1\%$) and highly significant for the financial sector. In line with our prior analyses, the actual value of the test statistic is small, suggesting that adding an estimated factor only slightly improves the model. Interestingly, the maximum BCV R^2 statistic is sometimes *negative* but nonetheless statistically significant for the healthcare sector (matching our prior finding of statistical significance in healthcare post-COVID). In other words, adding any of our extra estimated factors can lead to *worse* model performance in the healthcare sector, but under the null, we would expect to see an even more substantial reduction in model performance. This result aligns with prior theoretical analysis of PCA and factor models. For example, Paul (2007) showed that when testing whether a covariance matrix equals the identity, if the population maximum eigenvalue is only slightly larger than 1, the empirical maximum eigenvalue can often be statistically significant even while the empirical maximum eigenvector is approximately orthogonal to the population maximum eigenvector. Similarly, Owen and Wang (2016) emphasized that even if one can perfectly estimate a missing factor exposure, incorporating a missing factor exposure with a small effect size can lead to worse out-of-sample predictions by increasing the variance of the out-of-sample

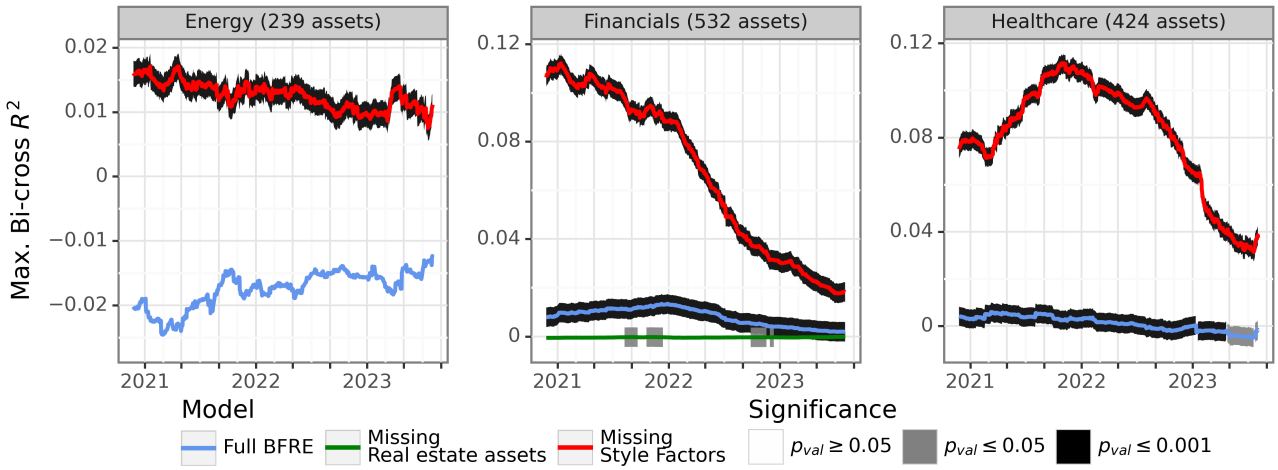


Figure 8: This figure plots the maximum bi-cross validation error statistic over time using a sliding window of 350 observations, both for the BFRE model as well as the ablation study where we remove the twelve style factors from the BFRE model. For the financial sector, we also perform an analysis where we remove the stocks classified as real estate assets. Using this test statistic, we also show the significance of the mosaic p-value every three months, as indicated by the black and gray shading. Note that occasionally, the p-value is significant even when the R^2 is negative, suggesting that the model does not fit perfectly but we do not know how to improve it (see Section 4.3 for discussion).

predictions. Overall, this analysis suggests that the unexplained correlations among healthcare assets are reasonably small.

We also continue our ablation test and repeat this analysis after removing the style factors from the BFRE model. Unsurprisingly, as shown by Figures 8, the maximum BCV R^2 values are substantially larger (often as high as 10%) and are uniformly highly significant in the ablation test. This gives more suggestive evidence that the BFRE model explains the most significant correlations among assets.

4.4 The missing factor among financial assets

In our previous analyses, we consistently found the most significant evidence against the null in the financial sector. We now investigate why this is the case. In short, we find that this result is driven by a set of assets in the real estate sector.

In particular, when running (approximate) sparse PCA as per Eq. 4.2 in the financial sector with $\ell = 20$ stocks based on data up to October 2020, 100% of the selected stocks have significant real estate exposure and 75% are primarily classified as "real estate" assets in the BFRE model. (In particular, the BFRE model has a "real estate" industry factor, and 75% of the selected assets have more exposure to the real estate industry factor than to any other industry factor.) To confirm this result, we recompute the maximum BCV R^2 statistic for the financial sector after removing real estate assets from the analysis. Figure 8 shows that we are unable to improve the model for the financial sector after removing the real estate assets.

Figure 9 also plots an out-of-sample estimate of the correlation matrix of the (estimated) residuals of the top 12 assets selected by the sparse PCA analyses with $\ell = 20$. The assets are ordered by the sign of the estimated maximum eigenvector, and the plot shows that the signs of the correlations among the assets match the signs of the maximum eigenvector. Namely, assets with the same sign are positively correlated, and assets with different signs are negatively correlated. This result gives

Figure 9: This figure shows the correlation matrix of the estimated residuals of the top twelve assets selected by the sparse PCA analysis in Eq. 4.2. Note that the sparse PCA is performed on data up to October 2020 and the correlation estimates are computed out-of-sample using data after October 2020. The assets are shown in order of the value of the estimated eigenvector $\hat{v}$, and the dotted black line shows the point at which the eigenvector's entries switch from negative to positive. As expected, we see largely positive correlations on the block diagonal outlined by the black lines and largely negative correlations otherwise.

additional confirmation that there are unexplained correlations among real estate assets and that incorporating this analysis could improve the BFRE model. That said, the correlation patterns in Figure 9 are relatively sparse—for example, the residuals of MGIC Investment, the Radian Group, and the Essent group (which all provide mortgage insurance) are highly correlated with each other, but they are only weakly correlated with the other selected real estate stocks. This supports our overall finding that the existing model appears to explain the most significant correlations among the assets.

5 Extensions

In this section, we give three additional extensions to the core methodology introduced in Section 3.

5.1 Adaptive choices of test statistic

In practice, it is not always obvious how to choose the most powerful test statistic from a set of candidate test statistics $S_1(\hat{\epsilon}), \ldots, S_d(\hat{\epsilon})$, which, for example, could each leverage different assumptions about the sparsity of any missing factor exposures. The mosaic permutation testing framework already gives the analyst many ways to answer this question, such as (i) setting the final statistic $S = \max_{i=1}^{d} S_i(\hat{\epsilon})$ to be the maximum of the candidates (as we do in our real application) or (ii) using a single test statistic which employs cross-validation to adaptively estimate the sparsity level. However, this section explores another particularly powerful and computationally efficient technique.

As a concrete example, consider the *quantile of maximum absolute correlations* (QMC) test statistic. Recall that $\widehat{\text{MaxCorr}}_j = \max_{j' \neq j} |\hat{C}_{j,j'}|$ is the maximum absolute estimated correlation between the residuals of asset j and another asset. The QMC statistic is the γ empirical quantile of the estimated

maximum absolute correlations:

$$S_\gamma(\hat{\epsilon}) = Q_\gamma\left(\left\{\widehat{\text{MaxCorr}}_j : j \in [p]\right\}\right). \tag{5.1}$$

The QMC statistic is the quantile variant of the mean maximum absolute correlation statistic from Section 1, and it is designed to increase power in sparse settings. For example, if only 1% of the residuals are correlated, the MMC statistic will be corrupted by noise, whereas the value of the QMC statistic with $\gamma = 0.99$ will depend primarily on the $\widehat{\text{MaxCorr}}_j$ values for non-null assets. However, how can we choose γ from a set of candidates $\gamma_1, \ldots, \gamma_d$? More generally, how can we choose among a set of d candidate statistics $S_1(\hat{\epsilon}), \ldots, S_d(\hat{\epsilon})$?

To do this, as notation, let $\{\tilde{\epsilon}^{(r)}\}_{r=1}^R$ denote mosaic permutations of $\hat{\epsilon}$ such that $\hat{\epsilon}$ and $\{\tilde{\epsilon}^{(r)}\}_{r=1}^R$ are all exchangeable as per Section 3. Consider an arbitrary function $f(\hat{\epsilon}, \tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)})$ which peeks at the mosaic permutations and aggregates evidence among all d test statistics $S_1, \ldots, S_d$. We refer to this function $f : \mathbb{R}^{(R+1)\times T\times p} \to \mathbb{R}$ as a "meta test-statistic." E.g., one could take f to be the maximum among all $i = 1, \ldots, d$ normalized differences between each candidate test statistic $S_i(\hat{\epsilon})$ and its mosaic permutations:

$$f(\hat{\epsilon}, \tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}) \overset{\text{e.g.}}{=} \max_{i=1}^{d} \frac{S_i(\hat{\epsilon}) - \sum_{r=1}^R S_i(\tilde{\epsilon}^{(r)})}{\sqrt{\widehat{\text{Var}}(\{S_i(\tilde{\epsilon}^{(r)})\}_{r=1}^R)}}. \tag{5.2}$$

To compute a p-value based on f, the key intuition is that $\{\tilde{\epsilon}^{(r)}\}_{r=0}^R$ are exchangeable, where for notational convenience we set $\tilde{\epsilon}^{(0)} := \hat{\epsilon}$. Therefore for any permutation $\pi : \{0, \ldots, R\} \to \{0, \ldots, R\}$, we have that

$$f(\tilde{\epsilon}^{(0)}, \tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}) \overset{\text{d}}{=} f(\tilde{\epsilon}^{(\pi(0))}, \tilde{\epsilon}^{(\pi(1))}, \ldots, \tilde{\epsilon}^{(\pi(R))}). \tag{5.3}$$

Thus, informally, we can compute a p-value based on any meta test-statistic by randomly permuting the order of $\{\tilde{\epsilon}^{(r)}\}_{r=0}^R$ and checking if this decreases the value of f. Algorithm 2 formally describes this procedure.

Algorithm 2 Adaptive meta-test statistic

Input: Returns $\mathbf{Y} \in \mathbb{R}^{T\times p}$, exposures $L_t \in \mathbb{R}^{p\times k}$ for $t \in [T]$, tiles $\{(B_m, G_m)\}_{m=1}^M$ and a meta test-statistic $f : \mathbb{R}^{(R+1)\times T\times p} \to \mathbb{R}$.

Step 1: Construct the mosaic residual estimate $\hat{\epsilon} \in \mathbb{R}^{T\times p}$ and its permuted variants $\tilde{\epsilon}^{(r)} \in \mathbb{R}^{T\times p}$ for $r = 1, \ldots, R$, as described in Algorithm 1. Set $\tilde{\epsilon}^{(0)} = \hat{\epsilon}$.

Step 2: Compute the original meta test-statistic $f(\tilde{\epsilon}^{(0)}, \ldots, \tilde{\epsilon}^{(R)})$.

Step 3: Sample uniformly random permutations $\pi_1, \ldots, \pi_K : \{0, \ldots, R\} \to \{0, \ldots, R\}$.

Step 4: Compute the final adaptive p-value

$$p_{\text{adaptive}} = \frac{1 + \sum_{\ell=1}^K \mathbb{I}\left(f(\tilde{\epsilon}^{(0)}, \tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}) \le f(\tilde{\epsilon}^{(\pi_\ell(0))}, \tilde{\epsilon}^{(\pi_\ell(1))}, \ldots, \tilde{\epsilon}^{(\pi_\ell(R))})\right)}{1 + K}. \tag{5.4}$$

Corollary 5.1. *p_{adaptive} is a valid p-value testing $\mathcal{H}_0$ under the same assumptions as Theorem 3.2.*

Note that Algorithm 2 uses *two* distinct layers of permutations—one set of mosaic permutations to construct $\{\tilde{\epsilon}^{(r)}\}_{r=1}^R$, and a second set of simple random permutations $\pi_1, \ldots, \pi_K$ to compute an adaptive p-value based on f. Although using the second layer of permutations adds conceptual complexity, the good news is it is extremely computationally efficient compared to cross-validating an expensive machine learning algorithm. As we shall see in Section 6, this procedure can be powerful and highly sparsity-adaptive.

5.2 Adaptively choosing the tiling

Although Section 3.3 gives a good default choice of tiling, another option is to *learn* a good choice of tiles that "separate" assets whose idiosyncratic returns are correlated. However, in general, if the tiling is chosen using $\mathbf{Y}$, $\hat{\epsilon}$ will not necessarily be invariant to any permutations under the null because of the dependence between $\{(B_m, G_m)\}_{m=1}^M$ and $\mathbf{Y}$. In other words, naive "double dipping" leads to inflated false positives.

However, we can sequentially choose the mth tile (B_m, G_m) based on the estimated returns from the previous $m-1$ tiles as long as our choice of (B_m, G_m) does not depend on the *order* of the rows within each of the previous $m-1$ tiles. Precisely, suppose that we can write B_m, G_m as functions b_m, g_m of the previous tiles as well as auxiliary randomness $U_m \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$:

$$B_m = b_m(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(m-1)}, U_m) \text{ and } G_m = g_m(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(m-1)}, U_m). \tag{5.5}$$

If b_m and g_m are invariant to permutations of the previous tiles, then p_{val} will be a valid p-value, as stated by the following lemma. Intuitively, the proof of the lemma follows from the fact that Eq. 3.10 still holds after conditioning on the tiles $\{(B_m, G_m)\}_{m=1}^M$ (see Appendix A for a proof).

Lemma 5.1. *Suppose for each $m = 1, \ldots, M$, (B_m, G_m) does not depend on the order of the rows within the first $m-1$ tiles. Formally, for any permutation matrices $P_j \in \mathbb{R}^{|B_j| \times |B_j|}$, we assume that*

$$b_m(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(m-1)}, U_m) = b_m(P_1\hat{\epsilon}_{(1)}, \ldots, P_{m-1}\hat{\epsilon}_{(m-1)}, U_m),$$

and the same holds when replacing b_m with g_m. Suppose also that Assumption 3.1 holds and, for simplicity, that $\{\epsilon_{t,j}\}_{t=1}^T \overset{\text{i.i.d.}}{\sim} P_j$ is i.i.d. for each $j \in [p]$. Then p_{val} is still a valid p-value testing $\mathcal{H}_0$ as in Theorem 3.2.

Remark 4. In Appendix A, we relax the i.i.d. assumption in Lemma 5.1 to a local exchangeability assumption (see Remark 6). For simplicity, we defer this extension to Appendix A.

Lemma 5.1 allows us to use many different methods to adaptively choose the tiling. For example, our default non-adaptive choice of tiling involved randomly partitioning the assets into D groups $[p] = G_{i,1} \cup \cdots \cup G_{i,D}$ for batch $i \in [I]$ of the observations. Instead, one could make an adaptive choice satisfying the permutation-invariance constraint from Lemma 5.1 as follows. For $i = 1$, we choose the groups randomly as before. Then, sequentially for $i \geq 2$, we let $\hat{\Sigma}^{(i)}$ be the empirical covariance estimator which (a) only uses information from the first $i-1$ batches and (b) only uses information from *within* tiles (and not between tiles). Precisely, fix any pair of assets $j, j' \in [p]$. Let $A = \{t \in [T] : (j, j') \in G_{i_0,d} \text{ for some } i_0 \leq i, d \in [D]\}$ be the set of time points in the first $i-1$ batches where assets j, j' are in the same group. (Note A depends on i, j, j' but for simplicity, we suppress this dependence). The within-tile covariance estimate is defined as

$$\hat{\Sigma}^{(i)}_{j,j'} = \frac{1}{|A|}\sum_{t \in A} \hat{\epsilon}_{t,j}\hat{\epsilon}_{t,j'} - \left(\frac{1}{|A|}\sum_{t \in A}\hat{\epsilon}_{t,j}\right)\left(\frac{1}{|A|}\sum_{t \in A}\hat{\epsilon}_{t,j'}\right).$$

It is easy to see that $\hat{\Sigma}^{(i)}$ does not depend on the ordering of the rows of any of the tiles. Thus, we can choose the groups for the ith batch based on $\hat{\Sigma}^{(i)}$ while preserving validity. In particular, we suggest choosing partitions which approximately solve the following optimization problem:

$$G_{i,1}, \ldots, G_{i,D} = \underset{G_{i,1},\ldots,G_{i,D} \text{ partition } [p]}{\arg\max} \sum_{j,j' \in [p]} |\hat{\Sigma}^{(i)}_{j,j'}| \cdot \mathbb{I}(j, j' \text{ are in different groups}). \tag{5.6}$$

In other words, for each i, we choose $G_{i,1}, \ldots, G_{i,D}$ to maximize the sum of the absolute estimated correlations between residuals that are not in the same group. Although exactly solving this optimization problem is computationally prohibitive, we can solve it approximately using a greedy randomized algorithm outlined in Appendix C.1.

5.3 Allowing the exposures to change with each observation

Motivated by our real applications, our analysis so far assumes that the exposures $L_t \in \mathbb{R}^{p\times k}$ are constant within tiles. However, if L_t changes with every observation, a naive application of our methodology would force the analyst to use tiles with only one observation, leading to zero power. A simple fix is to replace L_t with an *augmented* exposure matrix $L_t^\star$:

$$L_t^\star := \begin{cases} \begin{bmatrix} L_t & L_{t+1} \end{bmatrix} & t \text{ is odd} \\ \begin{bmatrix} L_{t-1} & L_t \end{bmatrix} & t \text{ is even} \end{cases} \in \mathbb{R}^{p\times 2k}. \tag{5.7}$$

Note that by construction, $L_t^\star$ only changes every two observations. For example, for the first two time points:

$$L_1^\star = L_2^\star = \begin{bmatrix} L_1 & L_2 \end{bmatrix} \in \mathbb{R}^{p\times 2k}.$$

Furthermore, since L_t is a submatrix of $L_t^\star$, if the null $\mathcal{H}_0$ holds for the original model $Y_t = L_t X_t + \epsilon_t$, it also holds for the augmented model $Y_t = L_t^\star X_t^\star + \epsilon_t$, since we can simply set the augmented factors $X_t^\star \in \mathbb{R}^{2k}$ to equal $(X_t, 0)$ for even t and $(0, X_t)$ for odd t. Thus, after augmenting the exposures, we can apply the mosaic permutation test (note that all tiles will contain exactly two observations to ensure $L_t^\star$ is constant within tiles). However, this does come at a cost, since we have to estimate twice as many nuisance parameters when estimating $\hat{\epsilon}$.

6 Do the mosaic residual estimates cause a loss of power?

In Section 1.1, we motivated the mosaic permutation test by asking how to check the statistical significance of a test statistic $S(\hat{\epsilon}^{\text{ols}})$ which was a function of the standard OLS residual estimates. To be precise, the mosaic permutation test does *not* allow one to do this—instead, it allows one to check the significance of a statistic $S(\hat{\epsilon})$ which is a function of *mosaic* residual estimates. Intuitively, we hope that the mosaic statistic is a good proxy for the OLS statistic, and indeed, we empirically see a high correlation between the mosaic and OLS test statistics in Figure 4. In this section, we analyze via simulations whether the use of the mosaic statistic leads to lower power than an oracle that uses the OLS statistic. Our simulations also show the effectiveness of the adaptive test statistic introduced in Section 5.1.

We conduct semisynthetic simulations where we set the exposures $L_t = L$ to be constant over time and equal to the real exposures from the BFRE model for the financial sector on April 17th, 2020. We sample the factors returns X_{tk} as i.i.d. standard Laplace random variables so that the factor returns have heavy tails. Similarly, we sample the residuals as follows:

$$\epsilon_t = \gamma_t + Z_t v \text{ for } \gamma_{t,j} \overset{\text{i.i.d.}}{\sim} \text{Laplace}, Z_t \overset{\text{i.i.d.}}{\sim} \text{Laplace and } v \in \mathbb{R}^p. \tag{6.1}$$

In words, the residuals are i.i.d. Laplace random variables plus an additional "missing" factor component, where $Z_t \in \mathbb{R}$ denotes the extra factor return at time t and $v \in \mathbb{R}^p$ denotes the corresponding

factor exposures. We choose v to be sparse, with $\lceil s_0 p \rceil$ nonzero coordinates chosen uniformly at random, and we set the nonzero coordinates to have value $\frac{\rho}{\sqrt{\lceil s_0 p \rceil}}$—thus, ρ measures the signal size and s_0 measures the sparsity. All simulations use $n = 200$ data points.

We compare the performance of three methods.

1. First, we apply the mosaic permutation test (MPT) with the default choice of tiling and the adaptive quantile maximum correlation (QMC) statistic from Eq. 5.1. In particular, since we do not know the optimal quantile value γ a priori, we use Eqs. 5.2 and 5.4 to compute an adaptive p-value which aggregates evidence across the test statistics using $\gamma \in \{0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99\}$.
2. Second, we compare this to the performance of the MPT with an "oracle" QMC statistic, where we pick the single value of γ which maximizes power—this is an oracle because, in a real analysis, the optimal value of γ would be unknown.
3. Lastly, we also compute the oracle QMC statistic applied to the OLS residuals $\hat{\epsilon}^{\text{ols}}$. We check the significance of the OLS oracle QMC statistic by comparing it to its true distribution under the null, with $\rho = 0$. We emphasize that this is a "doubly oracle" test statistic since in a real data analysis, we would not know the optimal choice of γ, nor would we know the the OLS statistic's null distribution. That said, comparing to this OLS statistic will help us understand whether the mosaic residual estimates cause a loss in power.

Figure 10 shows the results, namely that for various sparsities s_0 and signal sizes ρ, the MPT does not lose much power compared to either of the oracle tests. This suggests that our sparsity-adaptive QMC statistic effectively adapts to the unknown sparsity level. Furthermore, the MPT oracle and OLS double oracle—which use exactly the same test statistic but applied to different residual estimates—have reasonably similar power, suggesting that using the mosaic residuals does not lead to an unacceptable loss of power in this regime. This result should not be too surprising, since fundamentally $\hat{\epsilon}^{\text{ols}}$ and $\hat{\epsilon}$ are estimating the same residuals ϵ, and as a result, the OLS statistic $S(\hat{\epsilon}^{\text{ols}})$ should be extremely highly correlated with the mosaic statistic $S(\hat{\epsilon})$. Indeed, this correlation is exactly what we see empirically in our real application (Figure 4), where in all three sectors, the mosaic statistics are empirically at least $\geq 85\%$ correlated with the OLS statistics. Thus, in this simulation, we find that the mosaic permutation test is competitive even with an oracle method based on the OLS residuals.

7 Discussion

This paper introduces the mosaic permutation test, an exact and nonparametric goodness-of-fit test for factor models with known exposures. In an empirical application to the BlackRock Fundamental Equity Risk model, we demonstrate how to use the mosaic permutation test to diagnose and improve financial factor models. Additionally, our simulations and theory show the power and flexibility of the mosaic permutation test, which can be used in combination with a wide variety of test statistics to quickly detect unexplained correlations among variables. Lastly, although this paper focuses on applications to financial factor models, our methods can also be applied to test the goodness-of-fit of pre-existing factor models in any domain, including psychology (e.g., McCrae and John, 1992) and genetics (e.g., Gain and François, 2021).

That said, our work leaves open several possible directions for future research.

- Confidence intervals. This paper focuses on hypothesis testing, but it would be interesting to see if one could produce confidence intervals that quantify *how severely* the null is violated. Indeed, this

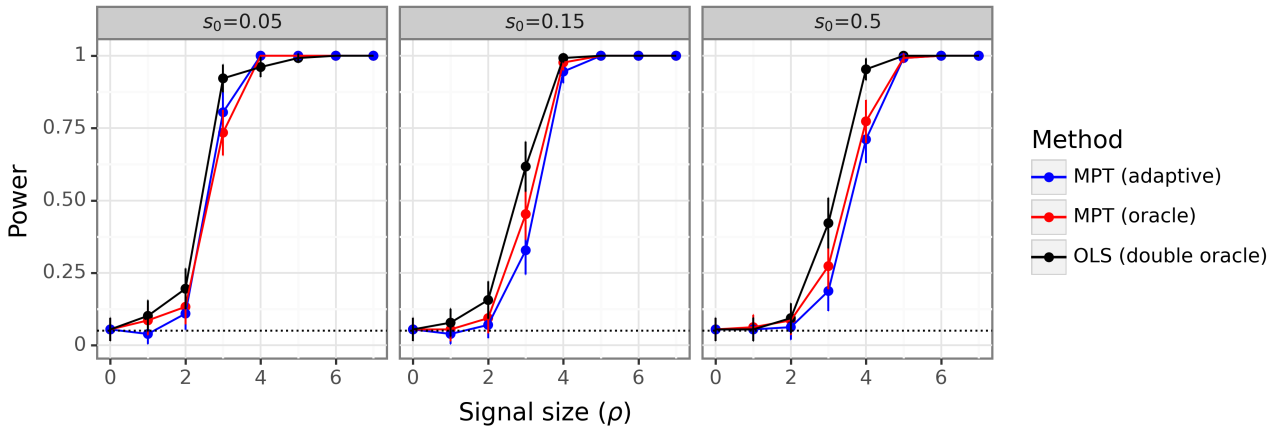


Figure 10: This figure shows the power of the mosaic permutation test with an adaptive QMC statistic as well as the power of the two oracles from Section 6. It shows that (1) the adaptive QMC statistic from Section 5.1 effectively adapts to the unknown sparsity of the alternative, and (2) the MPT does not lose much power compared to an oracle procedure using the OLS residuals $\hat{\epsilon}^{\text{ols}}$ in place of the mosaic residuals $\hat{\epsilon}$. The dotted black line shows the nominal level $\alpha = 0.05$. All methods control the false positive rate when $\rho = 0$ and $\mathcal{H}_0$ holds.

might help analysts understand the economic significance of any rejections made by our methods. We leave this promising question to future work.

- Anytime-valid tests. The mosaic permutation test p-values are not anytime-valid. This means that if one sequentially produces many mosaic p-values based on time series data, eventually, one will likely obtain a false positive by chance (Ramdas et al., 2023). This limitation should not affect our specific empirical findings, since the p-values in (e.g.) Figure 6 are so close to zero that they would remain significant even after an appropriate multiplicity correction. Nonetheless, producing an anytime-valid variant of the mosaic permutation test may be valuable for analysts who wish to monitor factor models over time.

- Factor models with known factor returns. This paper analyzes factor models with known exposures L_t. However, in some cases, it may be more realistic to assume the factor returns X_t are known. In future work, we extend the mosaic permutation test to apply to this new setting (among others). Yet this extension requires significant methodological innovations, so we defer it to a future paper (Spector et al., 2024).

- Using regularization: Our methods currently require the use of *unregularized* OLS regressions within each tile to estimate the residuals (see Remark 3). However, to increase power, it might be valuable to develop methodologies which can use *regularized* regressions instead.

- Robustness: It might be valuable to develop tests that are robust to slight inaccuracies in the exposures L_t. Indeed, this could also help relax the assumption that L_t is locally constant (Assumption 3.1), since small within-tile changes in L_t could be viewed as small "inaccuracies." Similarly, it would be useful to develop theory quantifying the robustness of the existing test, i.e., by bounding the excess error in some interpretable way.

8 Code and data availability

We implemented our methods in the python package `mosaicperm`. All other code used in the paper is available at `https://github.com/amspector100/mosaic_factor_paper/`. Although we are not able to make the BFRE model data available, we have provided a publicly available sample dataset that allows one to obtain qualitatively similar results (see the GitHub repository for more details).

References

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23):1806–1813.

Anderson, T. (2009). *An Introduction to Multivariate Statistical Analysis*. Wiley, wiley india pvt. limited. edition.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J., Duan, J., and Han, X. (2022). Likelihood ratio test for structural changes in factor models.

Bai, J. and Ng, S. (2006). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131(1):507–537.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends(R) in Econometrics*, 3(2):89–163.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer New York, NY.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):296–298.

Bender, J. and Nielsen, CFA, F. (2012). *The Fundamentals of Fundamental Factor Models*. John Wiley and Sons, Ltd.

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In *In A Festschrift for Erich L. Lehmann*, pages 28–48. Wadsworth Statistics.

Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):1–34.

Breitung, J. and Eickmeier, S. (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics*, 163(1):71–84. Factor Structures in Panel and Multivariate Time Series Data.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509 – 540.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2023). High-dimensional data bootstrap. *Annual Review of Statistics and Its Application*, 10(1):427–449.

D'Haultfœuille, X. and Tuvaandorj, P. (2023). A robust permutation test for subvector inference in linear regressions.

Dobriban, E. (2020). Permutation methods for factor analysis and PCA. *The Annals of Statistics*, 48(5):2824 – 2847.

Dobriban, E. and Owen, A. B. (2018). Deterministic Parallel Analysis: An Improved Method for Selecting Factors and Principal Components. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):163–183.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gain, C. and François, O. (2021). Lea 3: Factor models in population genetics and ecological genomics with r. *Molecular Ecology Resources*, 21(8):2738–2748.

Gonçalves, S. and Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495.

Grinold, R. and Kahn, R. N. (1994). Multiple-factor models for portfolio risk. *A practitioner's guide to factor models*, pages 59–85.

Guan, L. (2023). A conformal test of linear models via permutation-augmented regressions.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–185.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327.

Karoui, N. E. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *Journal of Machine Learning Research*, 19(5):1–66.

Kunsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3):1217 – 1241.

Lawley, D. N. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1/2):128–136.

Lei, L. and Bickel, P. J. (2020). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the svd and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564–594.

Owen, A. B. and Wang, J. (2016). Bi-Cross-Validation for Factor Analysis. *Statistical Science*, 31(1):119 – 139.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642.

Perry, P. O. (2009). Cross-validation for unsupervised learning.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference.

Romano, J. P. and Wolf, M. (2006). Improved nonparametric confidence intervals in time series regressions. *Journal of Nonparametric Statistics*, 18(2):199–214.

Rosenberg, B. and Marathe, V. (1976). Common factors in security returns: Microeconomic determinants and macroeconomic correlates. Research Program in Finance Working Papers 44, University of California at Berkeley.

Roy, S. N. (1953). On a Heuristic Method of Test Construction and its use in Multivariate Analysis. *The Annals of Mathematical Statistics*, 24(2):220 – 238.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Spector, A., Barber, R. F., and Candès, E. (2024+). Mosaic permutation tests for panel data.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Wen, K., Wang, T., and Wang, Y. (2023). Residual permutation test for high-dimensional regression coefficient testing.

A Proofs

A.1 Proof of Theorem 3.2 and Corollary 5.1

In this section, we prove Theorem 3.2 and Corollary 5.1. We do not explicitly prove Theorem 3.1 because it is a special case of Theorem 3.2 with $M = 2$ tiles.

Theorem 3.2. *Suppose Assumptions 3.1-3.2. Then under $\mathcal{H}_0$, Eq. 3.9 defines a valid p-value satisfying $\mathbb{P}(p_{\text{val}} \leq \alpha) \leq \alpha$ for any $\alpha \in (0, 1)$.*

Proof. The proof precedes in three steps. The first step reviews two useful consequences of Assumption 3.2 (namely Equations A.1 and A.2). The second step uses these results to show that the mosaic estimate $\hat{\epsilon}$ and its permuted variants $\tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}$ are jointly exchangeable. The final step shows the validity of the resulting p-value. Many of these steps use standard statistical arguments about exchangeability, but we include them all for completeness.

Step 1: Recall that $\epsilon_{(m)} \coloneqq \epsilon_{B_m, G_m} \in \mathbb{R}^{|B_m| \times |G_m|}$ denotes the residuals in the mth tile, for $m \in [M]$. Observe that under $\mathcal{H}_0$ and Assumption 3.2, for any permutation matrices $P_1 \in \mathbb{R}^{|B_1| \times |B_1|}, \ldots, P_M \in \mathbb{R}^{|B_M| \times |B_M|}$, we have the distributional equality:

$$(P_1 \epsilon_{(1)}, \ldots, P_M \epsilon_{(M)}) \overset{\text{d}}{=} (\epsilon_{(1)}, \ldots, \epsilon_{(M)}). \tag{A.1}$$

The above equation holds because each column (asset) in ϵ is separately permuted with a permutation that only swaps entries *within* tiles (not between tiles). Local exchangeability (Assumption 3.2) guarantees that this does not change the *marginal* distribution of each column of ϵ, and the null $\mathcal{H}_0$ guarantees that all columns of ϵ are independent.

Notably, Lemma A.3 uses this to prove the following result. Let $P_m^{(r)} \in \mathbb{R}^{|B_m| \times |B_m|}$ denote the randomly sampled permutation matrix for tile m for the $r = 1, \ldots, R$ randomizations in Algorithm 1, and let $P_m^{(0)} = I_{|B_m|}$ denote the identity matrix. Then the following holds:

$$\left[(P_1^{(r)}\epsilon_{(1)}, \ldots, P_M^{(r)}\epsilon_{(M)})\right]_{r=0}^{R} \text{ are exchangeable.} \tag{A.2}$$

This is a rather intuitive consequence of Equation A.1, but proving it requires dense notation, so to ease readability we defer the proof to Lemma A.3.

Step 2: As notation, recall from Equation 3.7 that the mth tile of the mosaic estimate $\hat{\epsilon}_{(m)}$ satisfies

$$\hat{\epsilon}_{(m)} \coloneqq \epsilon_{(m)} H_m$$

for a deterministic projection matrix $H_m \in \mathbb{R}^{|G_m| \times |G_m|}$. This notation implicitly uses the assumption that the exposures do not change within tiles (Assumption 3.1), as otherwise, H_m might change for different rows of $\hat{\epsilon}_{(m)}$.

Since H_m is a deterministic matrix for $m \in [M]$, Equation A.2 immediately implies that

$$\left[(P_1^{(r)}\epsilon_{(1)}H_1, \ldots, P_M^{(r)}\epsilon_{(M)}H_M)\right]_{r=0}^{R} \text{ are exchangeable.} \tag{A.3}$$

Recall that by definition, $P_m^{(0)}$ is the identity matrix, so $\hat{\epsilon}$ is simply a deterministic concatenation of $(P_1^{(0)}\epsilon_{(1)}H_1, \ldots, P_M^{(0)}\epsilon_{(M)}H_M)$. Similarly, by definition, $\tilde{\epsilon}^{(r)}$ is just the appropriate concatenation of $(P_1^{(r)}\epsilon_{(1)}H_1, \ldots, P_M^{(r)}\epsilon_{(M)}H_M)$. Thus, the previous equation implies that

$$(\hat{\epsilon}, \tilde{\epsilon}^{(1)}, \ldots, \tilde{\epsilon}^{(R)}) \text{ are exchangeable.} \tag{A.4}$$

Step 3: We now show that p_{val} is a valid p-value using Step 2. In particular, since $S : \mathbb{R}^{T \times p} \to \mathbb{R}$ is a deterministic function, we know that $(S(\hat{\epsilon}), S(\tilde{\epsilon}^{(1)}), \ldots, S(\tilde{\epsilon}^{(R)}))$ are exchangeable. This implies that if τ denotes the rank of $S(\hat{\epsilon})$ among $S(\hat{\epsilon}), S(\tilde{\epsilon}^{(1)}), \ldots, S(\tilde{\epsilon}^{(R)})$ where ties are broken uniformly at random and smaller ranks denote larger values, then $\tau \sim \text{Unif}(\{1, \ldots, R+1\})$. Note, however, that

$$p_{\text{val}} = \frac{\sum_{r=1}^{R+1} \mathbb{I}(S(\hat{\epsilon}) \leq S(\tilde{\epsilon}^{(r)})) + 1}{R+1} \geq \frac{\tau}{R+1},$$

where the deterministic equality follows because τ breaks ties uniformly at random but p_{val} is defined to always break ties conservatively. Thus, p_{val} stochastically dominates $\frac{\tau}{R+1}$, proving that $\mathbb{P}(p_{\text{val}} \leq \alpha) \leq \mathbb{P}(\tau \leq (R+1)\alpha) \leq \alpha$.

This also proves that if there are no ties with probability 1, then $p_{\text{val}} = \frac{\tau}{R+1} \sim \text{Unif}\left(\{\frac{1}{R+1}, \ldots, 1\}\right)$.

□

Corollary 5.1. p_{adaptive} *is a valid p-value testing* $\mathcal{H}_0$ *under the same assumptions as Theorem 3.2.*

Proof. Recall from Eq. A.4 that $\{\tilde{\epsilon}^{(r)}\}_{r=0}^{R}$ are exchangeable, where $\tilde{\epsilon}^{(0)} \coloneqq \hat{\epsilon}$. Also, recall that p_{adaptive} is defined as follows. Let $\pi_k : \{0, \dots, R\} \to \{0, \dots, R\}$ be uniformly random permutations for $1 \le k \le K$ and π_0 be the identity mapping. For a deterministic "meta test-statistic" $f : \mathbb{R}^{(R+1)\times T\times p} \to \mathbb{R}$, define $T_k = f(\tilde{\epsilon}^{(\pi_k(0))}, \dots, \tilde{\epsilon}^{(\pi_k(R))})$ for $0 \le k \le K$. Then

$$p_{\text{adaptive}} \coloneqq \frac{1 + \sum_{k=1}^{K} \mathbb{I}(T_0 \le T_k)}{K+1}.$$

Using the same logic as Step 3 in the proof of Theorem 3.2, it suffices to show that $\{T_k\}_{0\le k\le K}$ are exchangeable. However, since $\{\tilde{\epsilon}^{(r)}\}_{r=0}^{R}$ are exchangeable and $\pi_1, \dots, \pi_K$ are uniformly random permutations, $\{T_k\}_{0\le k\le K}$ are exchangeable as well. (This is a standard statistical argument, although for completeness we prove this result in Lemma A.2.) □

A.2 Proof of Lemma 5.1

We now prove a slightly more general version of Lemma 5.1. In particular, Lemma 5.1 assumed for simplicity that for each asset, the true residuals $\{\epsilon_{t,j}\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} P_j$ were i.i.d. This can be relaxed to a type of local exchangeability—however, we cannot directly assume Assumption 3.2, because Assumption 3.2 defines local exchangeability with respect to a *fixed* tiling, and in Lemma 5.1, the tiling $\{(B_m, G_m)\}_{m=1}^{M}$ is random. Instead, we must assume the following:

Assumption A.1. *Local exchangeability (Assumption 3.2) holds with respect to any fixed tiling $\{(\beta_m, \gamma_m)\}_{m=1}^{M}$ in the support of the learned (random) tiling $\{(B_m, G_m\}_{m=1}^{M}$. I.e., Assumption 3.2 holds for any fixed tiling satisfying $\mathbb{P}(\{(B_m, G_m)\}_{m=1}^{M} = \{(\beta_m, \gamma_m)\}_{m=1}^{M}) > 0$.*

Remark 5. Assumption A.1 always holds if the residuals $\{\epsilon_{t,j}\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} P_j$ are i.i.d. for each asset.

Remark 6. In practice, Assumption A.1 is not much stronger than Assumption 3.2. In particular, there is not much benefit to adaptively choosing the batches $B_1, \dots, B_M \subset [T]$—the main benefit is to choose the groups of assets $G_1, \dots, G_M \subset [p]$ adaptively to "separate" assets with highly correlated residuals. Thus, in Section 5.2, we suggested choosing $B_1, \dots, B_M$ so that each B_m is a member of a prespecified partition $\beta_1, \dots, \beta_{M^\star} \subset [T]$ of $[T]$, where $\beta_1 = \{1, \dots, 10\}, \beta_2 = \{11, \dots, 20\}$, etc. In this case, Assumption A.1 reduces to the the regular local exchangeability assumption (Assumption 3.2) with respect to the tiling $(\{\beta_m, [p]\}_{m=1}^{M^\star}$.

Having stated Assumption A.1, we now state and prove a more general variant of Lemma 5.1. As notation, recall that we assume that the mth batch $B_m = b_m(\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(m-1)}, U_m)$ and group $G_m = g_m(\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(m-1)}, U_m)$ are (potentially randomized) functions of the residuals from the first $m-1$ tiles, where $U_m \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ are independent uniform noise.

Lemma A.1. *Suppose Assumptions 3.1 and A.1 hold. Furthermore, for each $m = 1, \dots, M$, (B_m, G_m) are random but do not depend on the order of the rows within the first $m-1$ tiles. Formally, we assume that for any permutation matrices $P_j \in \mathbb{R}^{|B_j|\times|B_j|}$,*

$$b_m(\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(m-1)}, U_m) = b_m(P_1\hat{\epsilon}_{(1)}, \dots, P_{m-1}\hat{\epsilon}_{(m-1)}, U_m),$$

and the same holds when replacing b_m with g_m. Then p_{val} in Eq. 3.9 is a valid p-value testing $\mathcal{H}_0$.

Proof. It suffices to show that Equation A.1 holds conditionally on the choice of tiles $\{(B_m, G_m)\}_{m\in[M]}$:

$$(P_1\epsilon_{(1)}, \dots, P_M\epsilon_{(M)}) \overset{\text{d}}{=} (\epsilon_{(1)}, \dots, \epsilon_{(M)}) \mid \{(B_m, G_m)\}_{m\in[M]}. \tag{A.5}$$

After showing this, the original proof of Theorem 3.2 will go through after conditioning on the tiles.

The proof is simple but the notation is subtle. To ease comprehension, recall that by definition $\epsilon_{(m)} = \epsilon_{B_m,G_m}$ where B_m and G_m are random. In this proof, we will use the notation ϵ_{B_m,G_m} instead of $\epsilon_{(m)}$ to make the dependence on B_m and G_m explicit.

Let $\mathcal{T} = \{(B_m, G_m)\}_{m\in[M]}$ denote the choice of tiles and let $\tau = \{(\beta_m, \gamma_m)\}_{m\in[M]}$ denote an arbitrary *fixed* tiling in the support of $\mathcal{T}$. It suffices to show that for any fixed permutation matrices $P_1 \in \mathbb{R}^{|\beta_1|\times|\beta_1|}, \ldots, P_M \in \mathbb{R}^{|\beta_M|\times|\beta_M|}$,

$$(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M}) \overset{\mathrm{d}}{=} (\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M}) \mid \mathcal{T} = \tau. \tag{A.6}$$

To show this, we note that Assumption A.1 yields the *marginal* result that

$$(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M}) \overset{\mathrm{d}}{=} (\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M}). \tag{A.7}$$

To convert this to a conditional result, note that since $\mathcal{T}$ is a function of $\hat{\epsilon}$ (which itself is a deterministic function of ϵ), we can thus write $\mathcal{T}(\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M})$ as some function of $\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M}$. Using this fact, we define $\mathcal{T}_{\text{permute}} \coloneqq \mathcal{T}(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M})$ to be equal to the tiling we *would* have chosen based on the permuted residuals $(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M})$. Eq. A.7 now directly implies that

$$[(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M}), \mathbb{I}(\mathcal{T}_{\text{permute}} = \tau)] \overset{\mathrm{d}}{=} [(\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M}), \mathbb{I}(\mathcal{T} = \tau)]\,. \tag{A.8}$$

However, by assumption, $\mathcal{T}$ does not depend on the order of rows within each tile defined by $\{(B_m, G_m)\}_{m\in[M]}$. Thus, whenever $\mathcal{T} = \tau$, we have that $\{(\beta_m, \gamma_m)\}_{m\in[M]} = \{(B_m, G_m)\}_{m\in[M]}$ and thus $\mathcal{T} = \mathcal{T}_{\text{permute}}$, since $\mathcal{T}_{\text{permute}}$ is just the value of $\mathcal{T}$ after permuting the rows of each tile defined by $\{(\beta_m, \gamma_m)\}_{m\in[M]}$. Thus, $\mathcal{T} = \tau$ if and only if $\mathcal{T}_{\text{permute}} = \tau$, so $\mathbb{I}(\mathcal{T} = \tau) = \mathbb{I}(\mathcal{T}_{\text{permute}} = \tau)$. Combining this with the previous result yields that

$$[(P_1\epsilon_{\beta_1,\gamma_1}, \ldots, P_M\epsilon_{\beta_M,\gamma_M}), \mathbb{I}(\mathcal{T} = \tau)] \overset{\mathrm{d}}{=} [(\epsilon_{\beta_1,\gamma_1}, \ldots, \epsilon_{\beta_M,\gamma_M}), \mathbb{I}(\mathcal{T} = \tau)]\,. \tag{A.9}$$

This immediately implies that Eq. A.6 holds, since if $(X, Z) \overset{\mathrm{d}}{=} (Y, Z)$ for any random variables (X, Y, Z), then the conditional distributions $X \mid Z \overset{\mathrm{d}}{=} Y \mid Z$ must be equal as well. This concludes the proof.

□

A.3 Technical details

Lemma A.2. *Suppose $Z = (Z_1, \ldots, Z_n)$ are exchangeable random variables. Let $\pi_1, \ldots, \pi_K : [n] \to [n]$ be random permutations, sampled uniformly at random. Let $Y_k = (Z_{\pi_k(1)}, \ldots, Z_{\pi_k(n)})$ for $k \in [K]$ and set $Y_0 = (Z_1, \ldots, Z_n)$. Then $(Y_0, Y_1, \ldots, Y_K)$ are exchangeable.*

Proof. Fix any permutation $\tau : \{0, \ldots, K\} \to \{0, \ldots, K\}$. It suffices to show that

$$(Y_0, Y_1, \ldots, Y_K) \overset{\mathrm{d}}{=} (Y_{\tau(0)}, Y_{\tau(1)}, \ldots, Y_{\tau(K)}).$$

As notation, let P_k be the permutation matrix such that $P_k Z = Y_k$, for $k \in \{0, \ldots, K\}$. Let $J \coloneqq P_{\tau(0)}^{-1}$. The first step is to observe that

$$(Y_{\tau(0)}, Y_{\tau(1)}, \ldots, Y_{\tau(K)}) \coloneqq (P_{\tau(0)}Z, \ldots, P_{\tau(K)}Z) \overset{\mathrm{d}}{=} (P_{\tau(0)}JZ, \ldots, P_{\tau(K)}JZ),$$

where the distributional equality holds conditional on $\{P_k\}_{k\in[K]}$ since $Z \stackrel{\mathrm{d}}{=} JZ$ is exchangeable. The second step is to observe that by construction, $P_{\tau(0)}J = P_0$ is the identity permutation, and since $\{P_k\}_{k\in[K]}$ are uniformly random permutations, we know

$$(P_{\tau(0)}J, \dots, P_{\tau(K)}J) \stackrel{\mathrm{d}}{=} (P_0, \dots, P_K). \tag{A.10}$$

Combining the two previous results yields

$$(Y_{\tau(0)}, \dots, Y_{\tau(K)}) \stackrel{\mathrm{d}}{=} (P_{\tau(0)}JZ, \dots, P_{\tau(K)}JZ) \stackrel{\mathrm{d}}{=} (P_0Z, \dots, P_KZ) = (Y_0, \dots, Y_K)$$

where the second-to-last equality holds condition on Z. □

Lemma A.3. *Using the notation and assumptions from Theorem 3.2, the following holds:*

$$\left[(P_1^{(r)}\epsilon_{(1)}, \dots, P_M^{(r)}\epsilon_{(M)})\right]_{r=0}^{R} \text{ are exchangeable.}$$

Proof. To show this, we will show that for any fixed permutation $\pi : \{0, \dots, R\} \to \{0, \dots, R\}$,

$$\left[(P_1^{(r)}\epsilon_{(1)}, \dots, P_M^{(r)}\epsilon_{(M)})\right]_{r=0}^{R} \stackrel{\mathrm{d}}{=} \left[(P_1^{(\pi(r))}\epsilon_{(1)}, \dots, P_M^{(\pi(r))}\epsilon_{(M)})\right]_{r=0}^{R} \tag{A.11}$$

We will show the result in three steps.

Remark 7. Our proof essentially follows the proof of Lemma A.2, except we apply this argument simultaneously to each tile $m \in [M]$.

Step 1: Let Π_m denote the (random) inverse of $P_m^{\pi(0)}$. Equation A.1 (from Step 1 of the proof of Theorem 3.2) implies that

$$(\epsilon_{(1)}, \dots, \epsilon_{(M)}) \stackrel{\mathrm{d}}{=} (\Pi_1\epsilon_{(1)}, \dots, \Pi_M\epsilon_{(M)}), \tag{A.12}$$

where in particular, this holds conditional on $\Pi_1, \dots, \Pi_M$ since these are just permutation matrices, and thus it holds unconditionally as well.

Step 2: The previous observation implies that

$$\left[(P_1^{(\pi(r))}\epsilon_{(1)}, \dots, P_M^{(\pi(r))}\epsilon_{(M)})\right]_{r=0}^{R} \stackrel{\mathrm{d}}{=} \left[(P_1^{(\pi(r))}\Pi_1\epsilon_{(1)}, \dots, P_M^{(\pi(r))}\Pi_M\epsilon_{(M)})\right]_{r=0}^{R} \tag{A.13}$$

where the above equation holds conditional on all of the random permutation matrices $\{P_m^{(r)}\}_{m\in[M],r\in[R]}$; in particular, it follows by applying identical permutation matrices to both sides of Eq. A.12.

Step 3: Third, we observe that

$$\left[(P_1^{(r)}, \dots, P_M^{(r)})\right]_{r=0}^{R} \stackrel{\mathrm{d}}{=} \left[(P_1^{(\pi(r))}\Pi_1, \dots, P_M^{(\pi(r))}\Pi_M)\right]_{r=0}^{R} \tag{A.14}$$

To see this, it suffices to show $[P_m^{(r)}]_{r=0}^{R} \stackrel{\mathrm{d}}{=} [P_m^{(\pi(r))}\Pi_m]_{r=0}^{R}$ holds for a single fixed m, since the randomness for each $m \in [M]$ is completely independent. However, this fact about random permutations is exactly the content of Eq. A.10.

Combining steps two and three, we conclude:

$$\begin{aligned}\left[(P_1^{(\pi(r))}\epsilon_{(1)}, \dots, P_M^{(\pi(r))}\epsilon_{(M)})\right]_{r=0}^{R} &\stackrel{\mathrm{d}}{=} \left[(P_1^{(\pi(r))}\Pi_1\epsilon_{(1)}, \dots, P_M^{(\pi(r))}\Pi_M\epsilon_{(M)})\right]_{r=0}^{R} \\ &\stackrel{\mathrm{d}}{=} \left[(P_1^{(r)}\epsilon_{(1)}, \dots, P_M^{(r)}\epsilon_{(M)})\right]_{r=0}^{R}\end{aligned}$$

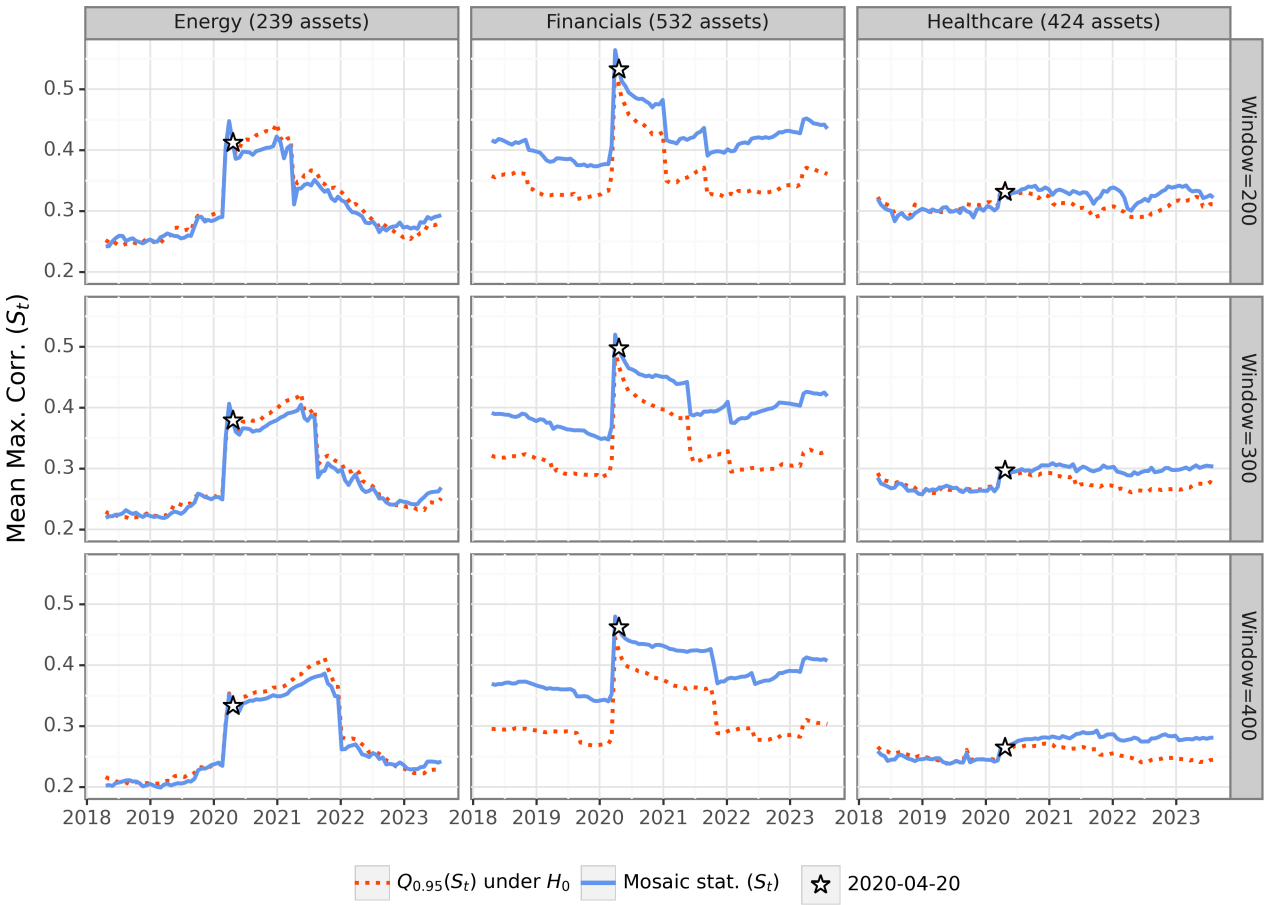


Figure 11: This figure exactly replicates the analysis from Figure 4, except that Figure 4 uses a window size of 350 and this plot varies the window size. In particular, for three industries, it shows the mosaic MMC statistic (plotted every two weeks) and its null quantile computed using different sliding window sizes. Note that the shape of the curves and the relative value of the statistic and its null quantile do not change substantially with different window sizes.

where in particular, the first line is a restatement of Step 2 and Step 3 proves that the second equality holds conditional on ϵ. This completes the proof.

□

B Additional empirical results

B.1 Sensitivity to the window size

Throughout the main text, we computed test statistics over time using a sliding window of size 350 observations. Although this is a somewhat arbitrary choice, we now show that varying the window size does not substantially change the results. In particular, Figure 11 replicates the analysis from Figure 4 and shows visually that the shapes of the curves plotting the mosaic test statistics and the mosaic null quantiles do not significantly change across different window sizes.

B.2 Results for additional sectors

We now show results for six additional sectors beyond the three from the main text: Consumer Discretionary, Consumer Staples, Industrials, Materials, Tech, and Utilities. In short, we find similar results to Section 4: either (i) we do not consistently reject the null at all or (ii) we can detect violations of the null $\mathcal{H}_0$ but the effect size is too small for us to consistently improve the model. The exception is the tech sector, where we can persistently improve the model (albeit by a small degree).

First, Figure 12 replicates the analysis from Figure 4 but for these additional sectors—in particular, it shows the mosaic MMC statistic computed in a sliding window of 350 observations for each sector as well as the 95% quantile of the permuted variants. We do not consistently reject the null in the utilities, materials, and consumer staples sectors, although we note that these sectors do not have many assets, so it is possible this is due to a lack of power. In the consumer discretionary, industrial, and tech sectors, we consistently reject the null.

Next, Figure 13 replicates the analysis from Figure 8 for six additional industries. I.e., it shows the value of the maximum bi-cross validation R^2 test statistic as described in Section 4.3 and the significance of this test statistic (computed using the mosaic permutation test). As in Section 4.3, in all sectors except one, we find that the test statistic is not consistently positive even when it is statistically significant. In other words, we may have power to detect violations of $\mathcal{H}_0$, but the effect size is sufficiently small that we cannot consistently improve the model (see Section 4.3 for more discussion of this phenomenon, which is predicted by several high-dimensional asymptotic theories). The only exception is the tech sector, where the maximum bi-cross R^2 is always positive. In contrast, if we perform this analysis after removing the style factors (as shown by Figure 13), the p-values become uniformly highly significant and the maximum bi-cross validation test statistics become larger. Overall, this supports the main finding from the main text, which is that the BFRE model does not fit perfectly but nonetheless explains the most significant correlations among asset returns.

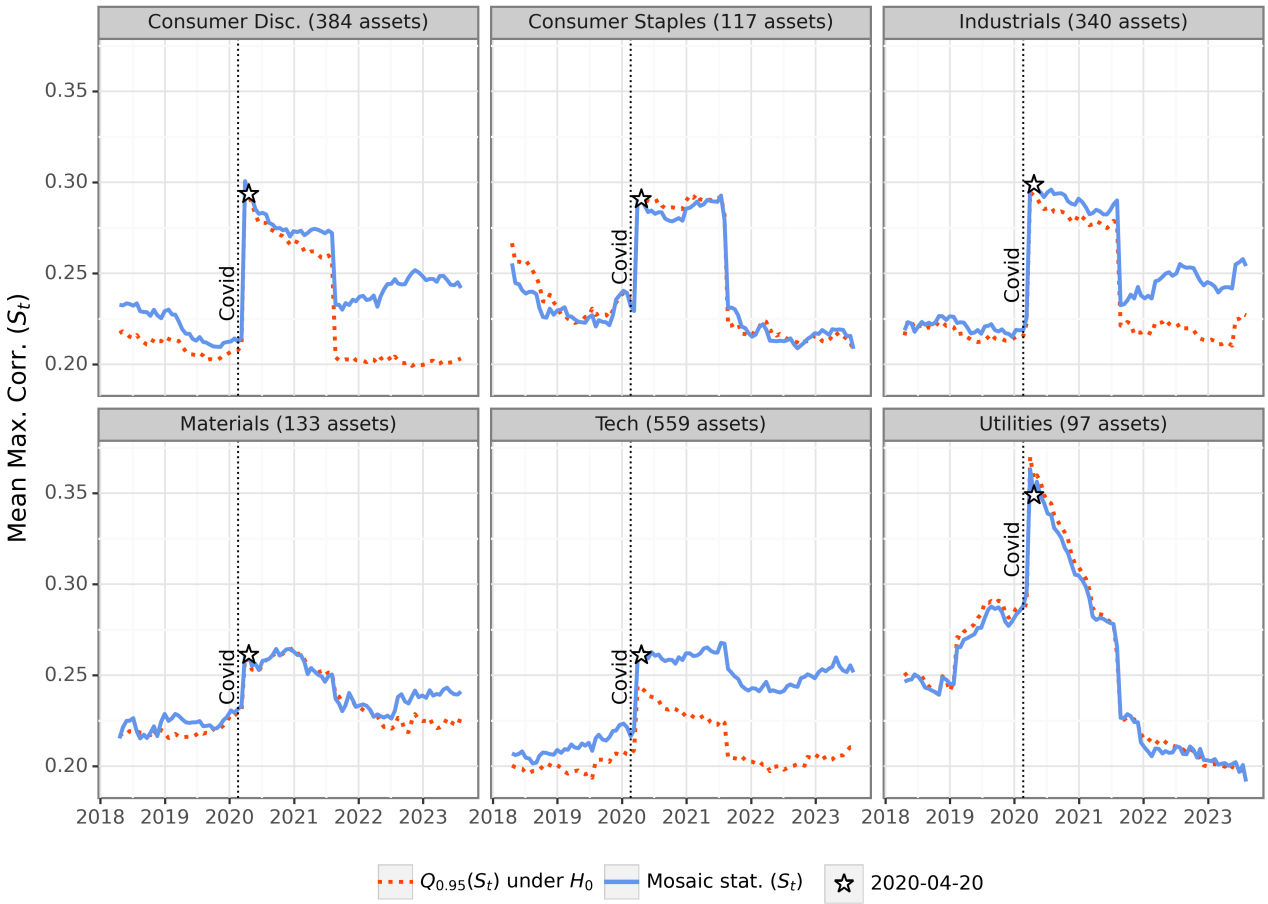


Figure 12: This figure replicates the analysis from Figure 4 but for six additional industries. In particular, for each industry, it shows the mosaic MMC statistic computed in a sliding window of 350 observations as well as significance threshold at the $\alpha = 0.05$ level.

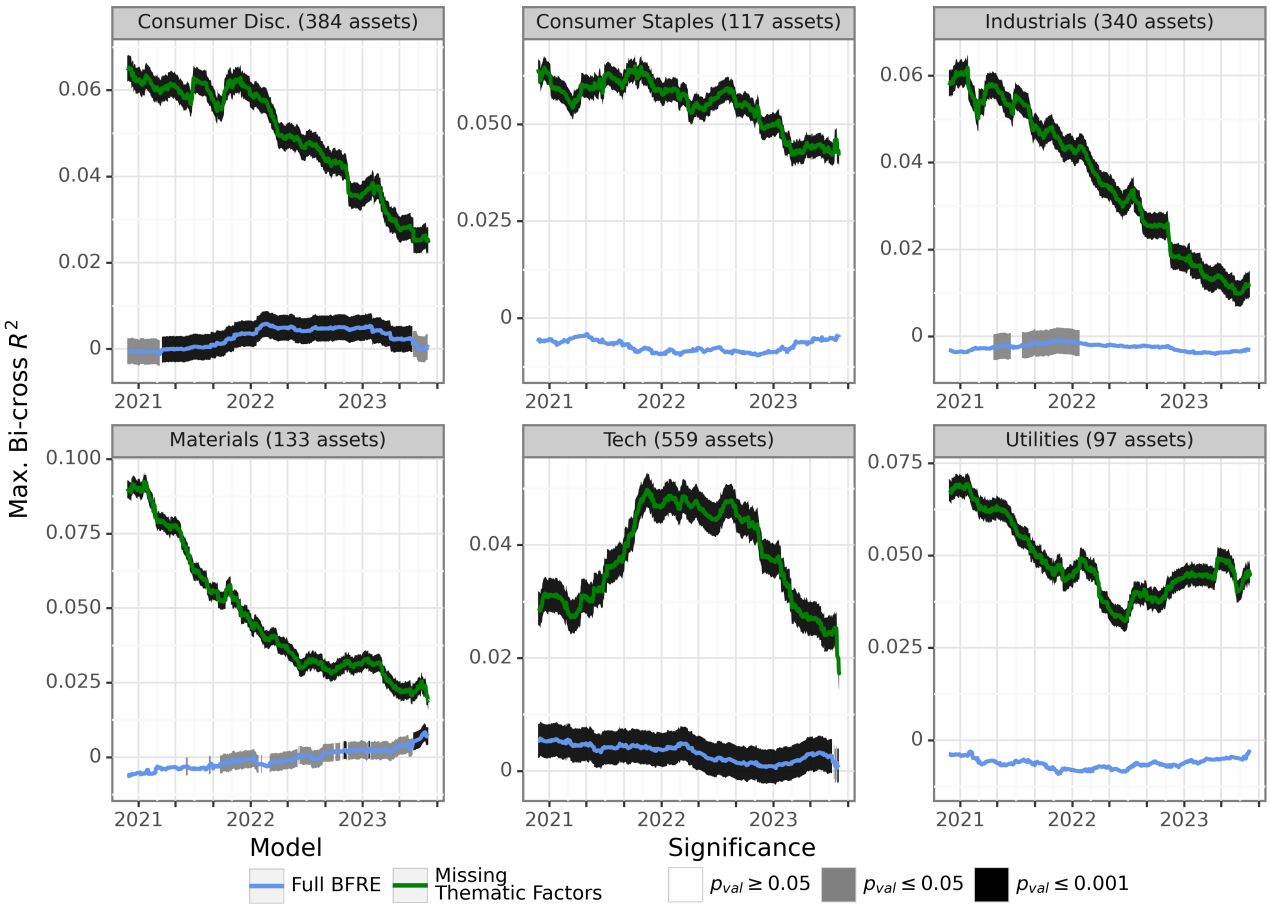


Figure 13: This figure replicates the analysis from Figure 8 for six additional industries, both for the full BFRE model and for an ablation study that removes the style factors from the BFRE model before performing the analysis. The shadings around each curve denote statistical significance.

C Additional methodological details

C.1 A greedy algorithm to adaptively choose good tiles

In Section 5.2, we introduced a general framework that allows the analyst to adaptively choose the tiling. One method we suggest involves approximately solving the following optimization problem:

$$G_1, \ldots, G_D = \underset{G_1,\ldots,G_D \text{ partition } [p]}{\arg\max} \sum_{j,j' \in [p]} |\hat{\Sigma}_{j,j'}| \cdot \mathbb{I}(j, j' \text{ are in different groups}). \tag{C.1}$$

Although it is unclear how to exactly solve this optimization problem, Algorithm 3 details an approximate greedy algorithm to solve Eq. C.1. In a nutshell, Algorithm 3 randomly initializes $G_1, \ldots, G_D$ to each contain one unique asset, and then iteratively adds each unassigned asset to the group which is minimally correlated with that asset. This is essentially a hierarchical "anti-clustering," since the goal is to *separate* highly correlated assets. Note that although this algorithm may not exactly maximize the objective in Eq. C.1, the p-value from Eq. 2.4 will still be valid.

Algorithm 3 Greedy algorithm to approximately solve Eq. C.1

Inputs: Estimated covariance matrix $\hat{\Sigma} \in \mathbb{R}^{p \times p}$, partition size $D \leq p$.

Step 1: Initialize $G_1 = \{j_1\}, G_2 = \{j_2\}, \ldots, G_D = \{j_D\}$ for unique randomly chosen assets $j_1, \ldots, j_D \in [p]$.

Step 2: While $G_1 \cup \cdots \cup G_D \neq [p]$ do the following:

- Randomly sample an element $j^\star \in [p] \setminus (G_1 \cup \cdots \cup G_D)$.
- Let $d^\star = \arg\min_{d \in [D]} \max_{j \in G_d} |\hat{\Sigma}_{j^\star,j}|$ denote the index of the group G_d which minimizes the maximum absolute estimated correlation between asset $j^\star$ and any asset $j \in G_d$.
- Reset $G_{d^\star} = G_{d^\star} \cup \{j^\star\}$.

Return: Partition $G_1, \ldots, G_D$.

C.2 Details for the sparse PCA algorithm

We now detail the greedy algorithm we used in Section 4.3 to approximately solve the sparse PCA problem:

$$\hat{v} \approx \max_{\|v\|_2 = 1} v^T \hat{C} v \text{ s.t. } \|v\|_0 \leq \ell. \tag{C.2}$$

In particular, let $\widehat{\text{MaxCorr}}_j := \max_{j' \neq j} |\hat{C}_{j,j'}|$ be the maximum absolute estimated correlation between asset j and asset j'. Let $S \subset [p]$ be the subset of indices of $[p]$ corresponding to the assets with the ℓ largest values of $\widehat{\text{MaxCorr}}_j$, so $|S| = \ell$. Finally, let $\hat{\nu}$ denote the top eigenvector of $\hat{C}_{S,S}$. We then define $\hat{v}$ as follows:

$$\hat{v}_j = \begin{cases} \hat{\nu}_j & j \in S \\ 0 & \text{else.} \end{cases}$$

In other words, S is the support of $\hat{v}$, and on S, $\hat{v}$ equals the maximum eigenvalue of $\hat{C}_{S,S}$. We picked this algorithm because it is conceptually simple and computationally cheap, but we have not explored other algorithms, which may have better performance. Indeed, the choice of sparse PCA algorithm is orthogonal to our contribution: the mosaic permutation test can be used in combination with a maximum bi-cross validation R^2 statistic based on any sparse PCA algorithm.

C.3 Additional details for the naive bootstrap Z-statistics

As discussed in Section 2.2, a common method to estimate the bias of a test statistic $S(\hat{\epsilon}^{\text{ols}})$ which estimates an underlying parameter θ is to compute a bootstrap bias estimate of the form

$$\widehat{\text{Bias}} = \frac{1}{B}\sum_{b=1}^{B} S(\hat{\epsilon}^{\text{ols},(b)}) - \theta_{\text{BS}},$$

where $\hat{\epsilon}^{\text{ols},(b)}$ is a bootstrapped residual matrix for $b = 1, \ldots, B$ and θ_{BS} is the true value of the parameter based on the empirical law of the data.

Recall that for the MMC test statistic used for the simulations in Figure 3, the parameter θ is the mean maximum absolute correlation of the correlation matrix $C^\star = \text{Cov}(\epsilon_1) \in \mathbb{R}^{p\times p}$ of the residuals. We can think of the bootstrap as a simulation where we sample each row of $\hat{\epsilon}^{\text{ols},(b)}$ i.i.d. from $\text{Unif}\left(\{\hat{\epsilon}_1^{\text{ols}}, \ldots, \hat{\epsilon}_T^{\text{ols}}\}\right)$. Thus, in the bootstrap "simulation," this means that θ_{BS} is equal to the mean maximum absolute correlation of the empirical correlation matrix of $\{\hat{\epsilon}_t^{\text{ols}}\}_{t\in[T]}$. In other words, we simply set $\theta_{\text{BS}} = S(\hat{\epsilon})$.