Factor models with many assets: strong factors, weak factors, and the two-pass procedure

Stanislav Anatolyev[1] — CERGE-EI and NES

Anna Mikusheva[2] — MIT

January 2021

[1]Addresses: CERGE-EI, Politickych vězňů 7, 11121 Prague 1, Czech Republic; New Economic School, Skolkovskoye shosse, 45, Moscow, Russia, 121353; e-mails stanislav.anatolyev@cerge-ei.cz, sanatoly@nes.ru.

[2]Address: Department of Economics, M.I.T., 50 Memorial Drive, Building E52, Cambridge, MA, 02142, USA; e-mail amikushe@mit.edu.

Abstract

This paper re-examines the problem of estimating risk premia in unconditional linear factor pricing models. Typically, the data used in the empirical literature are characterized by weakness of some pricing factors, strong cross-sectional dependence in the errors, and (moderately) high cross-sectional dimensionality. Using an asymptotic framework where the number of assets/portfolios grows with the time span of the data while the risk exposures of weak factors are local-to-zero, we show that the conventional two-pass estimation procedure delivers inconsistent estimates of the risk premia. We propose a new estimation procedure based on sample-splitting instrumental variables regression. The proposed estimator of risk premia is robust to weak included factors and to the presence of strong unaccounted cross-sectional error dependence. We prove the consistency of the new estimator, establish asymptotically valid inferences using Wald statistics, verify performance of the new procedure in simulations, and revisit some empirical studies.

`Key words:` factor models, risk premium, two-pass procedure, strong and weak factors, dimension asymptotics.

`JEL classification codes:` C33, C38, C58, G12.

1 Introduction

Since the introduction of the CAPM by Sharpe (1964) and Linner (1965), linear factor pricing models have grown into a very popular sub-field in asset pricing. Harvey, Liu and Zhu (2016) list hundreds of papers that propose, justify and estimate various factor pricing models. A typical paper in this area proposes a small set of observed risk factors that price the assets, that is, the expected excess return on an asset is equal to the quantity of risk taken (measured as a normalized covariance of the returns with the risk factors, so called betas) times risk premia. The two most famous factor pricing models are the market-factor CAPM and the three-factor Fama and French (1993) model. Other pricing factors are the momentum factor (Jegadeesh and Titman, 1993), the consumption-to-wealth ratio 'cay' (Lettau and Ludvigson, 2001), the liquidity factor (Pástor and Stambaugh, 2003), and so on. In recent years, there has been a burst in econometrics research that suggests how to correct the baseline estimation and inference in the face of many factors (for example, Kozak, Nagel, and Santosh, 2018), or how to judiciously select factors from a big pool without jeopardizing correct inference (for example, Feng, Giglio and Xiu, 2020).

Traditionally, one estimates the model using what is commonly known as the two-pass estimation procedure (Fama and MacBeth, 1973; Shanken, 1992),[1] where at the first pass one estimates risk exposures (betas) for each asset, and then, at the second pass, those estimates are used as regressors to estimate the risk premia. Asymptotic justification of this procedure, however, relies on assumptions that often do not hold up in realistic circumstances. Two types of violations of the idealistic setting have been noted in previous literature.

The first problem is one of weak (but priced) observed factors. Recent papers by Kan and Zhang (1999), Kleibergen (2009), Bryzgalova (2016), Burnside (2016), and Gospodinov, Kan and Robotti (2017) all point out that risk exposures (or betas) to some observed factors tend to be small to such an extent that their estimation errors are of the same order of magnitude as the betas themselves. This observed phenomenon is very similar to the widely studied weak instrument problem.

The second violation is a strong cross-sectional dependence in error terms, which in many cases can be modeled as a factor structure unaccounted for ('missing'). For example, recent literature shows that mismeasurement of the true risk factors leads to weakness of the observed factors and strong cross-sectional dependence in the errors (Kleibergen and Zhan, 2015), which may result in all sorts of distortions in estimation and inference in theory and in their non-reliability in practice (Kan and Zhang, 1999; Andrews, 2005; Kleibergen, 2009).

Along with the combination of the problems of missing factors and small betas, we also consider one very important empirical feature of the typically employed datasets – the presence of a large number of assets or portfolios often comparable to the number of periods over which returns are observed. We consider an asymptotic framework where the number of assets/portfolios grows with

[1] Sometimes the two-pass procedure is referred to as the Fama-MacBeth procedure (Fama and MacBeth, 1973). See Cochrane (2001, section 12.3) on their numerical equivalence when betas are time invariant. The method for obtaining valid standard errors that account for the two step nature of the procedure is given in Shanken (1992).

its time-series dimension. Such dimension asymptotics is likely to provide a more accurate asymptotic approximation to the finite sample properties of estimators and tests. The many-asset asymptotic framework has been utilized previously by Gagliardini, Ossola and Scaillet (2016), Lettau and Pelger (2020) and Feng, Giglio and Xiu (2020).

We show that within a dimension asymptotic framework the presence of small betas leads to a failure of the classical two-pass procedure, while the additional presence of missing factors exacerbates this problem. We propose econometric procedures that are robust to both these thorny issues with factors – the weakness of observed factors and the presence of unobserved factors in the errors – and, in contrast to the remedies proposed elsewhere, are easily implementable using standard regression tools (in particular, instrumental variables regressions and two-stage least squares). The estimators we propose are consistent; moreover, using the variance estimators (the construction of which we describe), standard inference tools such as t- and Wald tests can be applied in a conventional way.

Our new estimation approach makes use of the idea of sample-splitting in order to create multiple estimates for loadings β_i and to correct for the first-step estimation error via an instrumental variables regression. The presence of an unobserved (missing) factor structure in the error terms creates strong cross-sectional dependence in the panel of returns, which is similar to the classical omitted-variables problem in the second pass of the two-pass procedure. In order to correct for this missing factor structure, we use sample splitting to create reasonable proxies for missing factors even in a setting where one cannot consistently estimate the missing factor structure. The sample-splitting idea has appeared in the econometrics literature before, in particular, in Angrist and Krueger (1992) and Dufour and Jasiak (2001).

We explore the quality of the two-pass procedure and compare its performance with that of sample-splitting based estimators and existing alternatives in simulations calibrated to match the monthly returns of the 100 Fama-French sorted portfolios. We applied several existing procedures to the estimation of the momentum risk premium using real data on Fama-French portfolios. The important feature here is that the momentum is a tradable factor and hence there is an alternative estimate of the risk premia – the sample average excess return on this factor. Thus, we have a natural benchmark when comparing estimators. Lewellen, Nagel and Shanken (2010) showed that "any (sufficiently large)[2] set of assets perfectly explains the cross-section of expected returns so long as the (tested)[3] assets are not asked to price themselves (that is, ...the risk premia are not required to equal their expected returns)." From that perspective, having an estimate of risk premia coming from the pricing model, such as our split-sample estimator, is important even for tradable factors, as it allows one to test the pricing model by comparing this estimate to the average excess return.

There is a growing number of alternative suggestions for how to correct statistical inferences for either weak observed factors or a missing factor structure. Kleibergen (2009) proposes the use of weak identification robust inference procedures to account for weak observed factors, however, it can

[2]Our addition in brackets.
[3]Our addition in brackets.

only be applied to a relatively small number of assets/portfolios. An extreme version of the weak factors phenomenon, known as irrelevant factors and studied by Bryzgalova (2016), Burnside (2016) and Gospodinov, Kan and Robotti (2017), occurs when some observed factors are assumed to have zero loadings. The solutions to the irrelevant factors problem proposed in the literature usually suggest dimension reduction techniques to detect the irrelevant factors with a proviso to eliminate these detected irrelevant factors from further analysis. However, applying detection and elimination methods to the weak observed factors would lead to invalid inferences and large biases in the estimates of the risk premia for the remaining factors. Jegadeesh, Noh, Pukthuanthong, Roll and Wang (2019) suggest, simultaneously and independently, the use of sample-splitting in factor models in order to fix the errors-in-variables bias. Their proposed estimator works only when there are no missing factors.

Giglio and Xiu (2020) solve the problem of the missing factor structure by first running the Principle Component Analysis (PCA) on excess returns, pricing principle components, and from that deriving the risk premia of observed factors. The method of Giglio and Xiu (2020) successfully eliminates strong missing factors, but assumes from the outset that all important pricing factors can be uncovered by PCA. This assumption is critical for the validity of their procedure and contradicts the empirical findings of Lettau and Pelger (2020), who demonstrate that the out-of-sample performance using weak factors in addition to those uncovered by PCA is appreciably better than that of a model that uses the PCA factors only. According to Lettau and Pelger (2020), "PCA-based factors often miss low volatility components with high Sharpe ratios, which is a crucial aspect in asset pricing."

The paper is organized as follows. Section 2 introduces notation, discusses the relevance of our asymptotic approach, and argues for the presence of a significant factor structure in the errors. It also explains the asymptotic failure of the classical two-pass procedure and provides detailed intuition as to why this occurs. We propose our 'four-split' estimation method in Section 3, describe what motivates it and explain why it works. In Section 4, we state a formal theorem on the consistency of the newly proposed four-split estimator and establish the asymptotic validity of a properly constructed Wald test using the four-split estimator. The comparison of the newly proposed estimator with the existing alternatives is done in simulations and with an empirical example in Section 5. Appendix A contains main proofs, while auxiliary proofs and additional results appear in a Supplemental Appendix available on one of the authors' web-site.[4]

A word on notation: $0_{l,m}$ stands for a zero matrix of size $l \times m$, I_m is an $m \times m$ identity matrix; for an $m \times l_A$ matrix A and an $m \times l_B$ matrix B, (A, B) stands for the $m \times (l_A + l_B)$ matrix one obtains by placing the initial matrices side-to-side. Given a square matrix A, we denote by $\mathrm{dg}(A)$ a diagonal matrix of the same size with the same elements on the diagonal as matrix A, by $\mathrm{tr}(A)$ its trace, and by $\min \mathrm{ev}(A)$ and $\max \mathrm{ev}(A)$ – its minimal and maximal eigenvalue.

[4]`https://pages.nes.ru/sanatoly/Papers/ManyFM.htm`

2 Formulation of the problem

The research in factor asset pricing modeling typically proposes a small set of observed risk factors described by a vector F_t of (usually) small dimension k_F. An asset or portfolio of assets i with excess return r_{it} has exposure to several risk factors, which is quantified by the asset's betas $\beta_i = \text{var}(F_t)^{-1}\text{cov}(F_t, r_{it})$. A typical claim put forth in the linear factor-pricing theory is that exposure to risk (betas) fully determines the assets' expected excess returns. Particularly, there exists a k_F-dimensional vector of risk premia λ such that $Er_{it} = \lambda'\beta_i$.

From an econometric perspective, a correctly-specified linear factor-pricing model is equivalent to the following formulation:

$$r_{it} = \lambda'\beta_i + (F_t - EF_t)'\beta_i + \varepsilon_{it}, \tag{1}$$

where unobserved random error terms ε_{it} have mean zero and are uncorrelated with F_t. Here the statement $E\varepsilon_{it} = 0$ is equivalent to $Er_{it} = \lambda'\beta_i$, while uncorrelatedness between ε_{it} and F_t results from the definition of β_i. We treat λ and β_i as unknown parameters, while r_{it}, F_t, and ε_{it} are random variables.

Two-pass procedure. The estimation and inferences on risk prices, λ, are traditionally accomplished by a procedure known as the two-pass procedure (Fama and MacBeth, 1973; Shanken, 1992), applied to a data set consisting of a panel of asset excess returns $\{r_{it},\ i = 1, ..., N,\ t = 1, ..., T\}$ and observations of realized factors $\{F_t,\ t = 1, ..., T\}$. In the first step, one estimates β_i by running a time series OLS regression of r_{it} on a constant and F_t for each $i = 1, ..., N$. The second step produces an estimate of λ (denote it $\widehat{\lambda}_{TP}$) by regressing the time-average excess return $\frac{1}{T}\sum_{t=1}^{T} r_{it}$ on the first-step estimates, $\widehat{\beta}_i$. Under suitable conditions, $\widehat{\lambda}_{TP}$ is proved to be both consistent and asymptotically Gaussian. Discussions of the statistical properties of the two-pass procedure appear in Fama and MacBeth (1973), Shanken (1992), and Chapter 12 of Cochrane (2001).

This paper deviates from the classical Fama-MacBeth setting in three respects, which we label as (i) weak observed factors, (ii) many assets and (iii) missing factor structure.

Weak observed factors. Recent work by several prominent researchers raises the concern that the two-pass procedure may provide misleading estimates of risk premia; see, for example, Kan and Zhang (1999), Kleibergen (2009), Bryzgalova (2016), Burnside (2016), Gospodinov, Kan and Robotti (2017). The reason for these erroneous inferences is related to the empirical observation that either some column of $\beta = (\beta_1, ..., \beta_N)'$ is close to zero, or, more generally, the $N \times k_F$ matrix β appears close to one of reduced rank (less than k_F) for many well-known linear factor pricing models. According to Lettau and Pelger (2020), weak factors are empirically important for good performance of pricing models. They constructed factors that are impossible to uncover by PCA (i.e. weak factors) with the Sharpe ratio twice as high as those uncovered by PCA, and the out-of-sample pricing errors from a model that uses these weak factors are sizably smaller than those from a model that uses only strong factors.

Bryzgalova (2016), Burnside (2016), and Gospodinov, Kan and Robotti (2017) all recently developed improved inference procedures when some factors are completely *irrelevant* for pricing, that is, when true β_i are exactly zeros. Unfortunately, these procedures fail when the β_i are not zeros, but are small. A more empirically relevant case, which is in line with Lettau and Pelger (2020), resembles the widely studied weak instrument problem (Staiger and Stock, 1998): if some of the observed factors F_t are only weakly correlated with all the returns in the data set, then the noise that arises in the first-pass estimates of the corresponding components of β_i will dominate the signal, and the second-pass estimate of the risk premia λ will be oversensitive to small perturbations in the sample. In order to model the observed phenomenon, Kleibergen (2009) considered a drifting-parameter framework in which some component of β_i is modeled to be of order $O(\frac{1}{\sqrt{T}})$ assuming that the number of time periods, T, increases to infinity, while the number of assets, N, stays fixed. In such a setting the first-pass estimation error is of order of magnitude $O_p(\frac{1}{\sqrt{T}})$, which is comparable to the size of the coefficients themselves. This framework implies inconsistency of the two-pass estimates for the risk premium on small components, poor coverage of regular confidence sets even for the risk premium of strong factors, and asymptotic invalidity of classical specification tests and tests about risk premia.

Following this tradition, we also make use of drifting-parameter modeling. We assume that the $k_F \times 1$ vector of factors F_t can be divided into two subvectors: a $k_1 \times 1$ dimensional vector $F_{t,1}$ and a $k_2 \times 1$ vector $F_{2,t}$ (here $k_F = k_1 + k_2$) such that the risk exposure $\beta_{i,1}$ to factor $F_{t,1}$ is strong, while the risk exposure coefficients $\beta_{2,i}$ to factor $F_{2,t}$ jointly drift to zero at the rate $\sqrt{T}$. We make these order assumptions for risk exposures more accurate in the next section. A more general treatment of the near-degenerate rank condition considers some k_2-dimensional linear combination of factors (unknown to the researcher) to have a local-to-zero (of order $O(\frac{1}{\sqrt{T}})$) exposure coefficient, while the exposure to risk formed by the orthogonal k_1-dimensional linear combination remains fixed. All our results are easily generalizable to this setting, as we do not assume that the researcher knows which factors (or combination of factors) bear small coefficients of exposure. However, to simplify the exposition we stick to the division of factors into two sub-vectors.

One may argue that non-zero pricing on weak factors contradicts the Arbitrage Pricing Theory of Chamberlain and Rothschild (1983) that suggests that only strong factors are non-diversifiable and carry risk premia. In the face of empirical evidence showing that weak factors do carry risk premia and are important for factor pricing (e.g, Lettau and Pelger (2020) discussed above), the way to reconcile empirical evidence with the theory is to interpret weak factor modeling as an econometric device helpful in producing better finite-sample approximations. This device allows one to properly account for the uncertainty in the first step estimation rather than hide it under the rug of first step consistency. It is also worth pointing out that whether an observed factor is weak or strong depends, at least partially, on the data set at hand, its size and representativeness.

Many assets. In theoretical justifications of the two-step procedure (Shanken, 1992; Cochrane, 2001) it is common to assume that the number of assets, N, is fixed, while the number of periods T grows to infinity. We notice that in many common data sets that researchers use, the number of assets is large when compared to the number of time periods. The celebrated Fama-French data set provides returns on $N = 25$ sorted portfolios for about $T = 200$ periods. The often-used Jagannathan-Wang data set (Jagannathan and Wang, 1996) contains observations on $N = 100$ portfolios observed for $T = 330$ periods. Lettau and Ludvigson (2001) use Fama-French $N = 25$ portfolios, the returns for which are observed over $T = 141$ quarters. Gagliardini, Ossola and Scaillet (2016) use $N = 44$ industry portfolios observed during $T = 546$ months. In these cases it is hard to believe that the asymptotic results derived under the assumption that N is fixed would provide an accurate approximation of finite-sample distributions. Indeed, among other things, Kleibergen (2009) discovers that the bias of the two-pass estimate of risk premia is strongly and positively related to the number of assets if the total factor strength is kept constant.

In this paper we consider asymptotics when both N and T increase to infinity without restricting the relative speed. Recent papers by Kim and Skoulakis (2018) and Raponi, Robotti and Zaffaroni (2020) consider a factor pricing model in an asymptotic setting with $N \to \infty$ while T remains fixed and show inconsistency of the two-pass procedures. We can show that the procedures we propose are consistent in the setting with $N \to \infty$, fixed T for *ex-post risk premia* if we impose slightly stronger assumptions on the cross-sectional dependence of the error terms than the ones introduced in Assumption ERRORS below.

Missing factor structure. This paper deviates from the existing literature in the explicit acknowledgment of high cross-sectional dependence among the error terms ε_{it} in model (1). We assume that there exists, unknown and unobserved to the researcher, a factor v_t and loadings μ_i such that

$$\varepsilon_{it} = v_t'\mu_i + e_{it},$$

where the 'clean' errors e_{it} are only weakly cross-sectionally dependent to the extent that asymptotically we may ignore that dependence (the exact formulation of this assumption appears later). The assumptions on loadings μ_i guarantee that the factor structure is strong enough to be both detected empirically and asymptotically important for inferences. An insightful discussion of a weak versus strong factor structure and cross-sectional dependence can be found in Onatski (2012).

Kleibergen and Zhan (2015) provide numerous pieces of empirical evidence that residuals from many well-known estimated linear factor-pricing models have non-trivial factor structures. For example, they point out that the first three principle components of the residuals from different pricing-model specifications used in the seminal paper by Lettau and Ludvigson (2001) explain from 82% to 95% of all residual variation. The largest eigenvalue of the covariance matrix of residuals in all these examples is very large and strongly separated from the other eigenvalues. Combining this evidence with the theoretical results on the limiting distribution of eigenvalues from Onatski (2012), one would suspect there is at least one strong factor present in the residuals. At least five other prominent factor-pricing studies cited in Kleibergen and Zhan (2015) demonstrate similar evidence of strong factor structures not accounted for in the residuals.

Relation between factor structure and correct specification. One may wonder whether the fact that the errors ε_{it} in model (1) have a factor structure implies that the pricing model is misspecified. The answer is "no"; the linear factor pricing model describes the expectations of excess returns, while the factor structure in the errors is related to their covariances or co-movements. It is easy to see that if the risk exposure and risk premia on the variables F_t price the assets, then the variables F_t co-move the assets' returns and produce factor-structure dependence in the returns. However, not all co-movements of returns must carry non-zero risk premia; those co-movements can be placed in the error term without causing misspecification of the pricing model. The correct specification of a pricing model requires keeping those pricing factors $F_{t,2}$ that carry small coefficients of exposure $\beta_{2,t}$ and produce only a weak factor structure in the returns. Dropping such observed factors from the specification leads to asymptotically misleading inferences for the two-pass procedure.

Tradable factors. The literature on factor pricing distinguishes cases of tradable and non-tradable factors. If a specific factor F_t is a tradable portfolio and is supposed to be priced by the same pricing model, then $\lambda = EF_t$, and one can get an estimate of risk premia as the sample average of excess returns. We do not make this assumption and allow λ to differ from EF_t. However, even for tradable factors there is a value from having an alternative estimator based on the pricing equation (1). Lewellen, Nagel and Shanken (2010) showed that it is relatively easy to price the market with high cross-sectional R^2 by any set of portfolios as long as their number is large enough, but only if one does not enforce that the risk premia be equal to the average return. The cross-sectional R^2 of the pricing model is much smaller if one enforces such a restriction. Thus, having an estimator of the risk premia that does not use the $\lambda = EF_t$ condition and comparing it to the average excess return for tradable factors is a valuable test of the pricing model.

2.1 Assumptions on factor structure

We consider the problem of estimation and inference on the risk premia λ based on observations of returns $\{r_{it},\ i = 1, ..., N,\ t = 1, ..., T\}$ and factors $\{F_t,\ t = 1, ..., T\}$ obeying a correctly-specified factor-pricing model:

$$r_{it} = \lambda'\beta_i + (F_t - EF_t)'\beta_i + v_t'\mu_i + e_{it}, \tag{2}$$

where the random unobserved factor v_t has zero mean and is uncorrelated with F_t. The idiosyncratic error terms e_{it} also have zero mean and are uncorrelated with F_t and v_t. Let $\gamma_i' = \left(\beta_{1i}', \sqrt{T}\beta_{2i}', \mu_i'\right)$ and $\Gamma_N' = (\gamma_1, ..., \gamma_N)$ be the $k \times N$ matrix, where $k = k_F + k_v$. Technically, $\gamma_{i,N,T}$ is more accurate indexing, as parameters γ_i may change with the sample size as do all other features of the data generating process, but we drop N, T to reduce clutter.

Assumption FACTORS. The $k_F \times 1$ vector of observed factors F_t is stationary with finite fourth moments and a full-rank covariance matrix Σ_F. The $k_v \times 1$ latent factors v_t satisfies the following:

$$\begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (F_t - EF_t) \\ \eta_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Sigma_F^{-1} \widetilde{F}_t v_t' \\ \eta_{v,T} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_t \end{pmatrix} \Rightarrow \begin{pmatrix} N(0, \Omega_F) \\ \eta \\ \eta_v \end{pmatrix},$$

where $\text{vec}(\eta) \sim N\left(0_{k_F k_v,1}, \Omega_{vF}\right)$, $\eta_v \sim N(0_{k_v,1}, I_{k_v})$ and $\widetilde{F}_t = F_t - \frac{1}{T}\sum_{s=1}^{T} F_s$.

Assumption LOADINGS. As both N and T increase to infinity, we have $N^{-1}\Gamma_N'\Gamma_N \to \Gamma$, where Γ is a positive definite $k \times k$ matrix. Also assume that $\max_{N,T} \frac{1}{N}\sum_{i=1}^{N} \|\gamma_i\|^4 < \infty$.

Additional assumptions are placed on the error term e_{it} in Section 4. Most of our results can be understood if one imagines e_{it} as independent both cross-sectionally and in the time series direction, and also independent from all factors. However, our results hold under much more general assumptions, an extensive discussion of which we postpone to Section 4.

In this paper we treat the loadings β_i and μ_i as unknown non-random vectors, the true values of which may change with the sample sizes N and T. Assumption LOADINGS characterizes the size of the loadings as the sample size increases. Notice that the loadings on the factors $F_{t,1}$ and v_t are treated differently than the loadings on $F_{t,2}$. Following Onatski (2012), we will refer to the former as "strong factors" and the latter as "weak factors." The cross-sectional average of squared loadings is closely connected to the explanatory power the factors exhibit in cross-sectional variation. The assumptions we make on the loadings $\beta_{i,1}$ and μ_i guarantee that the explanatory power of the factors $F_{t,1}$ and v_t dominates the noise from the idiosyncratic error terms.

The loadings $\beta_{i,2}$ are asymptotically of the same order of magnitude as $\beta_{i,1}$ divided by $\sqrt{T}$. Assumption LOADINGS enforces that the standard deviation of the first-step estimate $\widehat{\beta}_{i,2}$ be of the same order of magnitude as $\beta_{i,2}$ itself. The modeling assumption that makes $\beta_{i,2}$ drift to zero at the rate $\sqrt{T}$ is similar to assumptions made in Kleibergen (2009). When N and T grow proportionally, our assumption may be re-written in terms of statements that $\sum_{i=1}^{N} \beta_{i,2}\beta_{i,2}'$ converges to a constant matrix, which are common in the weak factor model literature. Our results, however, do not need to restrict the relative rate of increase in N and T. It is also important that the assumptions on loadings μ_i are such that the unobserved factor v_t in the error terms is strong. This is consistent with the empirical observations in Kleibergen and Zhan (2015). This also guarantees that the presence of the factor structure plays an important role in the asymptotics of two-pass estimation.

2.2 Challenges of the two-pass procedure

As we formally show in Theorem 1 in Section 4, the two-pass procedure fails in the setting described above. Specifically, the two-pass estimate $\widehat{\lambda}_{TP,2}$ of the risk premia on weak factors $F_{t,2}$ is inconsistent and converges in probability to an incorrect value. The two-pass estimate $\widehat{\lambda}_{TP,1}$ of risk premia on strong factors $F_{t,1}$ is $\sqrt{T}$-consistent, but has a bias of order $\frac{1}{\sqrt{T}}$, the same order of magnitude as the standard deviation of its asymptotic distribution. This leads to invalid inferences on the risk premia. The failure of the two-pass estimator can be explained by an interplay of two biases that can be labeled as attenuation and omitted variable biases.

No missing factors case. The classical error-in-variables (attenuation) bias arises even when there is no missing factor structure as long as some factors are weak ($k_2 \geq 1$). For this paragraph only assume that there are no missing factors ($k_v = 0$). The first-pass estimate $\widehat{\beta}_i$ of the risk exposure coefficients β_i contains estimation errors that are stochastically of order $O_p(1/\sqrt{T})$:

$$\widehat{\beta}_i = \left(\sum_{t=1}^{T} \widetilde{F}_t \widetilde{F}_t' \right)^{-1} \sum_{t=1}^{T} \widetilde{F}_t r_{it} = (\beta_i + u_i)\big(1 + o_p(1)\big),$$

where $u_i = \frac{1}{T} \sum_{t=1}^{T} \Sigma_F^{-1} \widetilde{F}_t e_{it}$, and the $o_p(1)$ term is related to the difference between $\Sigma_F = E\big[(F_t - EF_t)(F_t - EF_t)'\big]$ and $T^{-1} \sum_t \widetilde{F}_t \widetilde{F}_t'$. As a result, the second-pass regression encounters an error-in-variables problem. In the case of exposure to a strong observed factor, the estimation error in $\widehat{\beta}_{i,1}$ is asymptotically negligible compared to the size of the coefficient $\beta_{i,1}$ itself, and so this estimation error does not jeopardize consistency. However, the estimation error in $\widehat{\beta}_{i,2}$ is asymptotically of the same order of magnitude as the coefficient itself. The first-pass estimation errors in $\widehat{\beta}_{i,2}$ behave like a classical measurement error as the estimation errors $u_{i,2}$ for different assets are asymptotically uncorrelated. This leads to a classical attenuation bias.

General case. In the presence of a strong factor structure in the errors ($k_v > 0$), the first-pass estimates have the following form:

$$\widehat{\beta}_i = \left(\sum_{t=1}^{T} \widetilde{F}_t \widetilde{F}_t' \right)^{-1} \sum_{t=1}^{T} \widetilde{F}_t r_{it} = \left(\beta_i + \frac{\eta_T \mu_i}{\sqrt{T}} + u_i \right) \big(1 + o_p(1)\big), \tag{3}$$

where $\eta_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Sigma_F^{-1} \widetilde{F}_t v_t' \Rightarrow \eta$. Again, the estimation error in $\widehat{\beta}_{i,1}$ turns out to be asymptotically negligible when compared to the sizes of risk exposures $\beta_{i,1}$ themselves, while the estimation errors in $\widehat{\beta}_{i,2}$ – which are now equal to $\eta_T \mu_i / \sqrt{T} + u_i$ – are of the size $O_p(1/\sqrt{T})$, which is the same order of magnitude as the $\beta_{i,2}$'s themselves. The estimation errors of $\widehat{\beta}_{i,2}$ distort the asymptotics and invalidate classical inferences. However, this time (when $k_v > 0$) the estimation errors do not behave like classical measurement errors in two respects. First, the estimation errors for different assets are correlated due to the presence of the common component η_T. Second, unless μ_i is cross-sectionally uncorrelated with β_i, the estimation error will be correlated with its own regressor β_i.

There is an additional issue classically known as an omitted variable bias. Let us look at the second pass (normalized) 'ideal' regression, which one can obtain by time-averaging equation (2):

$$\sqrt{T} \overline{r}_i = \sqrt{T} \widetilde{\lambda}' \beta_i + \eta_{v,T}' \mu_i + \sqrt{T} \overline{e}_i, \tag{4}$$

where $\eta_{v,T} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_t \Rightarrow \eta_v \sim N(0_{k_v,1}, I_{k_v})$. Here we introduced normalization $\sqrt{T}$ to make regression (4) more compatible with the classical OLS setup. The regression error terms $\sqrt{T}\overline{e}_i$ all have orders of magnitude of $O_p(1)$, zero means and finite variances. Imagine for a moment that we know β_i and μ_i for all assets. Then, regression (4) will take the form of a classic OLS regression, with regressors $\sqrt{T}\beta_{i,2}$ and μ_i being of order of magnitude $O(1)$, in the sense expressed in Assumption LOADINGS, that in the classical regression setting would lead to a $\sqrt{N}$-consistent and asymptotically Gaussian OLS estimator of the coefficients on them. The regressor $\sqrt{T}\beta_{i,1}$ is, in contrast, of order $O(\sqrt{T})$ and carries a lot of information which, in the classical regression setting, leads to an OLS estimator of the coefficient λ_1 on this regressor that is both super-consistent and asymptotically centered Gaussian. However, because μ_i is unobserved, it becomes a part of the error term in the second-pass regression. If $\Gamma_{\beta\mu} \neq 0_{k_F,k_v}$, then even if there were no first-pass estimation error and we knew β_i, running an OLS in a regression of $\sqrt{T}\overline{r}_i$ on $\sqrt{T}\beta_i$ would produce invalid results due to the omission of μ_i.

3 Sample-split estimator of the risk premia

3.1 Idea of the proposed solution

The case of no factor structure in the error terms. We begin by solving the easier case when no unobserved factor structure is present in the errors, while some observed factors are weak. In such a case we have a classical measurement error-in-variables problem, and we solve it by constructing a proper instrument with the help of a sample-splitting technique.

Let us divide the set of time indexes $t = 1, ..., T$ into two non-intersecting equal subsets T_1 and T_2. A natural choice is to make T_1 the first half of the sample, and T_2 its second half. Let us run the first-step regression twice – separately on each sub-sample:

$$\widehat{\beta}_i^{(j)} = \left(\sum_{t\in T_j} \widetilde{F}_t^{(j)} \widetilde{F}_t^{(j)\prime}\right)^{-1} \sum_{t\in T_j} \widetilde{F}_t^{(j)} r_{it} = (\beta_i + u_i^{(j)})\big(1 + o_p(1)\big)$$

for $j = 1, 2$, where $\widetilde{F}_t^{(j)} = F_t - \frac{1}{|T_j|}\sum_{t\in T_j} F_t$, $u_i^{(j)} = \frac{1}{|T_j|}\sum_{t\in T_j} \Sigma_F^{-1}\widetilde{F}_t^{(j)} e_{it}$, and the $o_p(1)$ term is related to the difference between Σ_F and $\frac{1}{|T_j|}\sum_{t\in T_j} \widetilde{F}_t^{(j)}\widetilde{F}_t^{(j)\prime}$.

We impose assumptions on the errors e_{it} which guarantee that the two sets of mistakes, $\{u_i^{(1)}, i = 1, ..., N\}$ and $\{u_i^{(2)}, i = 1, ..., N\}$, are conditionally independent. One may use an estimate of β_i from one sub-sample (for example, $\widehat{\beta}_i^{(1)}$) as a regressor while the other (in this example, $\widehat{\beta}_i^{(2)}$) as an instrument. This leads to a valid IV regression. Indeed, the second-step regression we run is

$$\overline{r}_i = \widetilde{\lambda}'\widehat{\beta}_i^{(1)} + \big(\overline{e}_i - \widetilde{\lambda}' u_i^{(1)}\big).$$

In this regression the regressor and the instrument are correlated since they both contain β_i, hence the relevance condition holds.

Similar ideas, such as sample splitting and jackknife-type estimators, have been previously employed in the many instruments literature (e.g., Hansen, Hausman and Newey, 2008). There, the number of instruments grows to infinity with the sample size, and the authors introduce a modeling assumption that makes the estimation error of the reduced-form coefficients be of the same order of magnitude as the coefficients themselves. This is parallel to the dimension asymptotics for a number of portfolios and the local-to-zero asymptotics for risk exposures of weak factors in our setup. In the many instrument setting, the regular TSLS estimator has a significant bias, and classical inferences are asymptotically invalid. Some proposed solutions employ the second-stage instrumental variables regression where, for each observation, the regressor is obtained from a first-stage regression run on a sub-sample that does not include that observation, and the original instrument is still used as an instrument (see Angrist, Imbens and Krueger (1999) and Dufour and Jasiak (2001)). This forces the first-stage error in the projection to be uncorrelated with the instrument for this specific observation. Sample-splitting or leave-one-out type procedures restore consistency and classical inferences. One can re-write the two pass-procedure as a GMM moment condition, then the two-pass corresponds to an IV estimator with many instruments in the framework of Newey and Windmeijer (2009). However, the problem of missing factor structure does not fall within the Newey and Windmeijer (2009) framework. The main departure is that Newey and Windmeijer (2009) consider i.i.d. sampling, while the observations in our model (indexed by both i and t) are highly dependent.

Raponi, Robotti and Zaffaroni (2017) and Kim and Skoulakis (2018) consider a similar phenomenon by assuming that $N \to \infty$ while T is fixed. These papers suggest correcting attenuation bias by directly estimating it. However, these estimation techniques would fail in the presence of missing factors in the error term.

The case of factor structure in the error terms. The model with an unobserved factor structure has an additional problem – the presence of omitted (and unobserved) variable μ_i in regression (4). Formula (3) suggests that we can obtain a noisy proxy for μ_i by taking the difference between two estimates for the same β_i obtained from different sub-samples. Consider two non-intersecting subsets of time indexes, T_1 and T_2, and assume they have the same number, say τ, of time indexes. Then

$$\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)} = \frac{\eta_\tau^{(1)} - \eta_\tau^{(2)}}{\sqrt{\tau}} \mu_i + \big(u_i^{(1)} - u_i^{(2)}\big).$$

Notice that both the coefficient on μ_i and the noise term $u_i^{(1)} - u_i^{(2)}$ are of the same order of magnitude $O_p(1/\sqrt{\tau})$. This means that neither the signal dominates the noise – and thus we need a correction to account for the noise, – nor the noise dominates the signal, and thus the proxy is not useless.

Assume that $k_v \leq k_F$, which implies that we have a larger number of proxies than needed, and we have a choice among them. Now we assume that we have a fixed and full-rank $k_v \times k_F$ matrix A, and use $A\big(\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}\big)$ as the proxy. The idea is to regress the average return $\overline{r}_i$ on $\widehat{\beta}_i^{(1)}$ and $A\big(\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}\big)$ instead of on unobserved β_i and μ_i. This solves the omitted-variables part of the problem, but the error-in-variables issue still remains. That problem we solve via instrumental variables upon additional sample splitting. The ultimate idea goes as follows: split the sample into four equal sub-samples along the time dimension; calculate the first-pass estimates of risk exposures for all four sub-samples; run an instrumental variables regression using $\widehat{\beta}_i^{(1)}$ and $A\big(\widehat{\beta}_i^{(1)}-\widehat{\beta}_i^{(2)}\big)$ as regressors and $\widehat{\beta}_i^{(3)}$ and $\big(\widehat{\beta}_i^{(3)}-\widehat{\beta}_i^{(4)}\big)$ as instruments. We may repeat the procedure by circulating sub-sample indexes to improve efficiency.

3.2 Algorithm for constructing four-split estimator

Divide the set of time indexes into four equal non-intersecting subsets T_j, $j=1,...,4$.

(1) For each asset i and each subset j run a time-series regression to estimate the coefficients of risk exposure:

$$\widehat{\beta}_i^{(j)}=\left(\sum_{t\in T_j}\widetilde{F}_t^{(j)}\widetilde{F}_t^{(j)\prime}\right)^{-1}\sum_{t\in T_j}\widetilde{F}_t^{(j)}r_{it}.$$

(2) Run an IV regression of $\overline{r}_i=\frac{1}{T}\sum_{t=1}^T r_{it}$ on regressors $x_i^{(1)}=\big(\widehat{\beta}_i^{(1)\prime},(\widehat{\beta}_i^{(1)}-\widehat{\beta}_i^{(2)})'A_1'\big)'$ with instruments $z_i^{(1)}=\big(\widehat{\beta}_i^{(3)\prime},(\widehat{\beta}_i^{(3)}-\widehat{\beta}_i^{(4)})'\big)'$, where A_1 is a non-random $k_v\times k_F$ matrix of rank k_v. Let $\widehat{\lambda}^{(1)}$ be the TSLS estimate of the coefficient on regressor $\widehat{\beta}_i^{(1)}$.

(3) Repeat step (2) three more times exchanging indexes 1 to 4 circularly; that is, the 2^{nd} regression is an IV regression of $\overline{r}_i$ on regressors $x_i^{(2)}=\big(\widehat{\beta}_i^{(2)\prime},(\widehat{\beta}_i^{(2)}-\widehat{\beta}_i^{(3)})'A_2'\big)'$ with instruments $z_i^{(2)}=\big(\widehat{\beta}_i^{(4)\prime},(\widehat{\beta}_i^{(4)}-\widehat{\beta}_i^{(1)})'\big)'$; denote the estimate as $\widehat{\lambda}^{(2)}$, etc.

(4) Obtain the four-split estimate as $\widehat{\lambda}_{4S}=\frac{1}{4}\sum_{j=1}^4\widehat{\lambda}^{(j)}$.

(5) In order to compute an estimate of the covariance matrix for $\widehat{\lambda}_{4S}$, denote by $X^{(j)}$ the $N\times k$ matrix of stacked regressors used in the j^{th} IV regression, and by $Z^{(j)}$ the $N\times k_z$ matrix of instruments from this regression (here $k_z=2k_F$ and $k=k_F+k_v$). Let $P_Z=Z\left(Z'Z\right)^{-1}Z'$, calculate the following matrices:

$$\widehat{\Sigma}_0=\frac{1}{N}\sum_{i=1}^N\begin{pmatrix}\widetilde{z}_i^{(1)}\widehat{\epsilon}_i^{(1)}\\ \dots\\ \widetilde{z}_i^{(4)}\widehat{\epsilon}_i^{(4)}\end{pmatrix}\begin{pmatrix}\widetilde{z}_i^{(1)}\widehat{\epsilon}_i^{(1)}\\ \dots\\ \widetilde{z}_i^{(4)}\widehat{\epsilon}_i^{(4)}\end{pmatrix}',\quad G=\begin{pmatrix}G_1&0_{k,k}&0_{k,k}&0_{k,k}\\0_{k,k}&G_2&0_{k,k}&0_{k,k}\\0_{k,k}&0_{k,k}&G_3&0_{k,k}\\0_{k,k}&0_{k,k}&0_{k,k}&G_4\end{pmatrix},$$

where $\widehat{\epsilon}_i^{(j)}$ is i^{th} residual from the $(j)^{th}$ IV regression, $\widetilde{z}_i^{(j)}=X^{(j)\prime}Z^{(j)}(Z^{(j)\prime}Z^{(j)})^{-1}z_i^{(j)}$, and $G_j=\frac{1}{N}X^{(j)\prime}P_{Z^{(j)}}X^{(j)}$. Denote $R=(1,1,1,1)'\otimes\left(\frac{1}{4}I_{k_F}\;0_{k_F,k_v}\right)'$, a $4k\times k_F$ matrix. Then,

$$\widehat{\Sigma}_{4S}=\frac{1}{N}R'G^{-1}\widehat{\Sigma}_0G^{-1}R+\frac{1}{T}\widehat{\Omega}_F,$$

where $\widehat{\Omega}_F$ is a consistent estimator of the long-run variance of F_t.

3.3 Comparison with existing approaches

Kleibergen (2009) introduced weak observed factors, showed their impact on the estimation of the risk premia and proposed robust procedures inspired by weak IV robust tests. An important prerequisite for the robust tests is consistent estimation of the $N \times N$ cross-sectional covariance matrix of the pricing errors. Kleibergen (2009) assumes that N is fixed, as consistent estimation of the covariance matrix becomes problematic with a growing number of assets. A recent paper by Kleibergen, Kong and Zhan (2020) demonstrates the problems arising from large N. The authors propose finite-sample inferences for large N and limited T that differ dramatically from asymptotic inferences valid for fixed N and $T \to \infty$. The finite-sample approach of Kleibergen, Kong and Zhan (2020) allows for an arbitrary covariance matrix but requires Gaussianity of the pricing errors. The four-split approach we propose has an asymptotic justification and does not place distributional assumptions on the errors. However, we restrict the complexity of the covariance matrix by imposing the assumption of a latent k_v-dimensional factor structure in the pricing errors. This allows us to create proxies for the missing factor loadings and to account for the biases that arise from the cross-sectional correlation of the pricing errors.

Giglio and Xiu (2020) propose a three-pass method to estimate the risk premium of an observed factor. The procedure extracts factors from a set of assets returns using the PCA, calculates the risk premia of the extracted factors and then linearly transforms the estimated risk premia to price the observed factor. The three-pass method is robust both to missing strong factors as well as to weak or even irrelevant observed factors. The important assumption behind the three-pass method is that all priced risk factors are pervasive and can be uncovered by the PCA, the assumption we do not require. From this perspective, we consider our method as complimentary to Giglio and Xiu (2020). Recently, Lettau and Pelger (2020) questioned the abovementioned assumption and demonstrated that the PCA often misses weak factors that are empirically important for pricing. Lettau and Pelger (2020) propose an alternative to the PCA that finds statistical factors explaining both the covariance matrix and the expected returns of the assets, but, in contrast to the papers discussed above, it does not consider the issue of inference on the risk premia.

4 Theoretical statements

4.1 Assumptions about idiosyncratic errors

Assumption ERRORS.

(i) The random vectors $e_t = (e_{1t}, ..., e_{Nt})'$ are serially independent conditionally on $\mathcal{F}$, and $E(e_t|\mathcal{F}) = 0$, where $\mathcal{F}$ the sigma-algebra generated by variables $(F_1, ..., F_T)$ and $(v_1, ..., v_T)$.

(ii) Let $\rho(t, s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_{it}e_{is}$. Assume $\sup_t \sup_{s \neq t} E\left[(1 + \|F_t\|^4)(\rho(s, t)^2 + 1)\right] < C$.

(iii) Let $S_t = \frac{1}{N} \sum_{i=1}^{N} e_{it}^2$. Assume $\frac{\sqrt{N}}{T} \sum_{t=1}^{T} \widetilde{F}_t S_t = o_p(1)$ and $\frac{1}{T} \sum_{t=1}^{T} \widetilde{F}_t \widetilde{F}_t' S_t \to^p \Sigma_{SF^2}$.

(iv) Let $W_t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \gamma_i e_{it}$. Assume $E\left[(1 + \|F_t\|^2)\|W_t\|^2\right] < \infty$.

Assumption ERRORS are high-level assumptions introduced to establish our results formally. The main goal is to allow for very flexible weak cross-sectional dependence among the idiosyncratic errors, as well as flexible conditional heteroscedasticity and dependence in higher-order moments of errors and factors. Serial independence of errors as stated in Assumption ERRORS(i) is consistent with the efficient market hypothesis and the unpredictability of asset returns, and is generally consistent with empirical evidence and tradition in the literature. This assumption may be weakened, though we do not pursue this in the current paper.

The random variables $\rho(s,t)$ stand for a (normalized) empirical analog of the error autocorrelation coefficient, S_t is an empirical variance, and W_t is a (normalized) weighted average error. These variables are normalized so that they are stochastically bounded when the errors are cross-sectionally i.i.d. In order to clarify the content of Assumption ERRORS and to show that our assumptions are more flexible than those typically made in the literature, first, we provide a set of more restrictive primitive assumptions that are common in the literature and that guarantee the validity of Assumption ERRORS. Second, we also provide an empirically relevant example not covered by the primitive assumptions but which satisfies our more general Assumption ERRORS.

Assumption ERRORS*

(i) The factors $\{F_t,\, t = 1, ..., T\}$ are independent from the errors $\{e_{it},\, i = 1, ..., N,\, t = 1, ..., T\}$; the error terms $e_t = (e_{1t}, ..., e_{Nt})'$ are serially independent and identically distributed for different t with $Ee_{it} = 0$ and $\sup_{i,t} Ee_{it}^4 < \infty$.

(ii) Let $\mathcal{E}_{N,T} = E\left[e_t e_t'\right]$ be the $N \times N$ cross-sectional covariance matrix. For some positive constants a, c and C, we have $\lim_{N,T} \frac{1}{N} \mathrm{tr}(\mathcal{E}_{N,T}) = a$ and

$$c < \liminf_{N,T\to\infty} \min \mathrm{ev}\left(\mathcal{E}_{N,T}\right) < \limsup_{N,T\to\infty} \max \mathrm{ev}\left(\mathcal{E}_{N,T}\right) < C.$$

(iii) $E\left|\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (e_{it}^2 - Ee_{it}^2)\right|^2 < C.$

Lemma 1 *Assumptions LOADINGS and ERRORS* imply Assumption ERRORS.*

The primitive Assumption ERRORS* is very close to those standard in the literature. Numerous papers that establish inferences in factor models commonly assume that the set of factors $\{F_t,\, t = 1, ..., T\}$ is independent from the set $\{e_{it},\, i = 1, ..., N,\, t = 1, ..., T\}$, though cross-sectional dependence of errors is allowed; see, for example, Assumption D in Bai and Ng (2006). Many papers allow for both time-series and cross-sectional error dependence. We exclude time-series dependence, which is justified by the efficient-market hypothesis in our application. Assumption ERRORS*(ii) is intended to impose only weak dependence cross-sectionally as expressed by the covariance matrix; similar assumptions appear in Onatski (2012) and in Bai and Ng (2006).

Our high-level Assumption ERRORS is more general than the more standard primitive Assumption ERRORS*. In particular, our assumptions allow for very flexible conditional heteroscedasticity in the error terms and time-varying cross-sectional dependence, which seems relevant when we consider observed factors that characterize market conditions like the momentum factor. Consider the following example.

Example 1. Assume that errors e_{it} have the following weak latent factor structure:

$$e_{it} = \pi_i' w_t + \eta_{it},$$

where (w_t, F_t) is stationary, w_t is $k_w \times 1$, serially independent conditional on $\mathcal{F}$, with $E(w_t|\mathcal{F}) = 0$ and $E(w_t w_t') = I_{k_w}$ (normalization). Assume $E\left[(\|F_t\|^4 + 1)(\|w_t\|^4 + 1)\right] < \infty$. We assume that the loadings satisfy the condition $\sum_{i=1}^N \pi_i \pi_i' \to \Gamma_\pi$ (the factors w_t are weak), and $N^{-1/2} \sum_{i=1}^N \pi_i \gamma_i' \to \Gamma_{\pi\gamma}$. Assume that the random variables η_{it} are independent both cross-sectionally and across time, are independent from w_t and F_t, and have mean zero and finite fourth moments and variances σ_i^2 that are bounded above and are such that $N^{-1} \sum_{i=1}^N \sigma_i^2 \to \sigma^2$. As proven in the Supplemental Appendix, this example satisfies Assumption ERRORS.

An interesting feature of this example is that it allows the errors to be weakly cross-sectionally dependent to the extent that they may possess a weak factor structure. Moreover, this factor structure may be closely related to the observed factors F_t, which causes the cross-sectional dependence among the errors e_{it} to change with the observed factors F_t and allows a very flexible form of conditional heteroskedasicity. Indeed, the conditional cross-sectional covariance is

$$E(e_{it} e_{jt}|\mathcal{F}) = \pi_i' E(w_t w_t'|\mathcal{F}) \pi_j + \mathbb{I}_{\{i=j\}} \sigma_i^2.$$

Since we do not restrict $E(w_t w_t'|\mathcal{F})$ beyond the proper moment conditions, the strength of any cross-sectional dependence as well as error variances may change stochastically depending on the realizations of the observed factors. This flexibility is extremely relevant for observed factors such as the momentum. For example, one may consider $w_t = \varsigma_t g(F_t, F_{t-1}, ...)$, where $\varsigma_t \sim N(0, 1)$ is independent from all other variables; then for a proper choice of the function $g(\cdot)$ one may observe higher volatility and cross-sectional dependence of the idiosyncratic error for higher values of the observed factor F_t.

4.2 Asymptotic properties of the two-pass procedure

Let us denote $\widetilde{\lambda} = \lambda + \frac{1}{T} \sum_{t=1}^T F_t - EF_t$, a random quantity known as ex-post risk premia. It was introduced in Shanken (1992). Now let us introduce two asymptotically important terms. The first term we refer to as "attenuation bias" is

$$B^A = -\left(\sum_{i=1}^N \widehat{\beta}_i \widehat{\beta}_i'\right)^{-1} \sum_{i=1}^N u_i u_i' \widetilde{\lambda},$$

while the second term we call the "omitted variable bias" is

$$B^{OV} = \left(\sum_{i=1}^N \widehat{\beta}_i \widehat{\beta}_i'\right)^{-1} \sum_{i=1}^N \widehat{\beta}_i \frac{\mu_i'}{\sqrt{T}} \big(\eta_{v,T} - \eta_T' \widetilde{\lambda}\big).$$

These terms are not biases in an exact sense as they are random, but rather they are sample analogues of the expressions that are classically known as attenuation and omitted variable biases. Notice that both quantities are infeasible as they depend on unobserved errors e_{it}, unobserved factors v_t and unknown parameters λ and μ_i. Both terms are $k_F \times 1$ vectors. Let B_1^A and B_1^{OV} denote $k_1 \times 1$ sub-vectors containing the first k_1 components, while B_2^A and B_2^{OV} are $k_2 \times 1$ sub-vectors of the last k_2 components of B^A and B^{OV}, correspondingly. We also adopt the following notation: $\Gamma_{\beta_2\mu}$ is the $k_2 \times k_\mu$ sub-block of matrix Γ (defined in Assumption LOADINGS) corresponding to the limit of $N^{-1}\sum_{i=1}^N \sqrt{T}\beta_{i,2}\mu_i$. Other sub-matrices are denoted similarly.

Theorem 1 *Assume that the samples $\{r_{it},\ i = 1, ..., N,\ t = 1, ..., T\}$ and $\{F_t,\ t = 1, ..., T\}$ come from a data-generating process that satisfies factor-pricing model (2) and assumptions FACTORS, LOADINGS and ERRORS. Let $\widehat{\lambda}_{TP}$ denote the estimate obtained via the conventional two-pass procedure. Let both N and T increase to infinity without restrictions on relative rates. Then the following asymptotic statements hold simultaneously:*

$$\begin{pmatrix} \sqrt{T} B_1^{OV} \\ B_2^{OV} \end{pmatrix} \Rightarrow \Big((I_{k_\beta}; \widetilde{\eta})\Gamma(I_{k_\beta}; \widetilde{\eta})' + \mathcal{I}_{k_2}\Sigma_u \mathcal{I}_{k_2}\Big)^{-1} (\Gamma_{\beta\mu} + \widetilde{\eta}\Gamma_{\mu\mu})(\eta_v - \eta'\lambda),$$

$$\begin{pmatrix} \sqrt{T} B_1^{A} \\ B_2^{A} \end{pmatrix} \Rightarrow -\big((I_{k_\beta}; \widetilde{\eta})\Gamma(I_{k_\beta}; \widetilde{\eta})' + \mathcal{I}_{k_2}\Sigma_u \mathcal{I}_{k_2}\big)^{-1} \mathcal{I}_{k_2}\Sigma_u \lambda,$$

$$\sqrt{T}\big(\widetilde{\lambda} - \lambda\big) \Rightarrow N(0, \Omega_F),$$

and

$$\begin{pmatrix} \sqrt{NT}\big(\widehat{\lambda}_{TP,1} - \widetilde{\lambda}_1 - B_1^A - B_1^{OV}\big) \\ \sqrt{N}\big(\widehat{\lambda}_{TP,2} - \widetilde{\lambda}_2 - B_2^A - B_2^{OV}\big) \end{pmatrix} = O_p(1),$$

where $\Sigma_u = \Sigma_F^{-1}\Sigma_{SF^2}\Sigma_F^{-1}$, $\mathcal{I}_{k_2} = \begin{pmatrix} 0_{k_1,k_1} & 0_{k_1,k_2} \\ 0_{k_2,k_1} & I_{k_2} \end{pmatrix}$ is a $k_F \times k_F$ matrix, and $\widetilde{\eta} = \mathcal{I}_{k_2}\eta$ is a $k_F \times k_v$ random matrix (with η as in Assumption FACTORS).

Theorem 1 states the rates of convergence for different parts of the two-pass estimator. Notice that the theorem does not impose a relative rate of increase between N and T. One observation is that the two-pass procedure cannot estimate λ at a rate faster than $\sqrt{T}$ despite the fact that the dataset has NT observations of portfolio excess returns, and one could expect the $\sqrt{NT}$ rate. This comes from the fact that the correct specification (1) if averaged across time, gives

$$\overline{r}_i = \widetilde{\lambda}\beta_i + \overline{\varepsilon}_i. \tag{5}$$

Thus, even if β_i were known, the 'true' coefficient $\widetilde{\lambda}$ in the only ideal regression we have (that is, regression of average return on β_i) differs from the parameter λ we want to estimate, by the term $\frac{1}{T}\sum_{t=1}^T F_t - EF_t$, which, if multiplied by $\sqrt{T}$, is asymptotically zero mean Gaussian with variance Ω_F. Notice also that if the limits of the normalized B^{OV} and B^A are non-zero, then these terms (together with $\widetilde{\lambda}_1$) asymptotically dominate estimation. The two-pass estimate $\widehat{\lambda}_{TP,2}$ of the risk premia on weak factors $F_{t,2}$ is inconsistent and, asymptotically, has a poorly-centered, non-standard distribution. The two-pass estimate $\widehat{\lambda}_{TP,1}$ of risk premia on strong factors $F_{t,1}$ is $\sqrt{T}$-consistent, but this estimate has a bias of order $\frac{1}{\sqrt{T}}$ and an asymptotically non-standard distribution as long as some of observed factors are weak ($k_2 > 0$). This makes standard inferences (based on the usual t-statistics) invalid.

4.3 Consistency of the four-split estimator

Theorem 2 *Assume that the samples $\{r_{it},\ i = 1, ..., N,\ t = 1, ...T\}$ and $\{F_t,\ t = 1, ..., T\}$ come from a data-generating process that satisfies factor pricing model (2) and Assumptions FACTORS, LOADINGS and ERRORS. Assume that $k_F \geq k_v$. Let both N and T increase to infinity, then*

$$\sqrt{T}(\widehat{\lambda}_{4S,1} - \lambda_1) = \sqrt{T}(\widetilde{\lambda}_1 - \lambda_1) + O_p(1/\sqrt{N}) \Rightarrow N(0, \Omega_F),$$

$$\sqrt{\min\{N,T\}}(\widehat{\lambda}_{4S,2} - \lambda_2) = O_p(1).$$

Theorem 2 establishes the consistency rate for the four-split estimator $\widehat{\lambda}_{4S}$ under exactly the same assumptions we showed the failure of the two-pass procedure. The four-split estimator for the risk premia on the strong observed factor is $\sqrt{T}$-consistent, asymptotically equivalent to $\widetilde{\lambda}_1$ and asymptotically Gaussian, while the four-split estimate of the risk premia on the weak observed factor is consistent, and the rate of convergence depends on the relative size of N and T. Theorem 2 shows that the four-split estimator has superior asymptotic properties in comparison to the classical two-pass procedure.

Theorem 2 can be generalized to allow local misspecification of the pricing model. The statement of the theorem remains true if the missing factors carry a risk premium of size up to $1/\sqrt{T}$. One way of expressing this is to assume that the expectation of v_t is not zero but is of order no larger than $1/\sqrt{T}$. The presence of the risk premia on the missing factors does not change the first stage estimation. The proof of Theorem 2 only uses that $\eta_{v,T} = \frac{1}{\sqrt{T}}\sum_{t=1}^T v_t$ is asymptotically $O_p(1)$ but not its mean-zero property.

If we consider an asymptotic setting where $N \to \infty$ while T remains fixed, one can prove that under slightly stronger assumptions on the error term, the four-split estimate is consistent for the ex-post risk premia $\widetilde{\lambda}$, and the t-statistics are asymptotically Gaussian. The distinction between ex-post and ex-ante risk premia is discussed in Shanken (1992) as well as in Raponi, Robotti and Zaffaroni (2017) and in Kim and Skoulakis (2018). We need to strengthen the cross-sectional dependence assumptions on the error terms to the extent that the law of large numbers and cental limit theorems hold when summation is done over the cross-sectional index only. The assumptions in Raponi, Robotti and Zaffaroni (2017) and Kim and Skoulakis (2018) are of this type.

One important assumption for the validity of our procedure is that we know the number of missing factors k_v. One may combine our estimator with a consistent selector of the number of factors as Onatski (2009), Bai and Ng (2002) or Gagliardini, Ossola and Scaillet (2019) do. A similar assumption/suggestion is used in Giglio and Xiu (2020) where properties of the estimator crucially depend on the proper choice of the number of factors. Underestimation of the number of missing factors leads to the omitted variable bias, while overestimation may produce weak instrument distortion on step (2) of our algorithm and calls for weak IV robust inferences.

4.4 Inference procedures using four-split estimator

Theorem 2 shows that the new four-split estimator is consistent but does not provide a basis for confidence set construction or testing. Apparently, the stated assumptions are not strong enough to obtain the asymptotic distribution of the four-split estimator. Below we formulate the needed additional high-level assumptions and establish a result about statistical inferences using the four-split estimator. We also provide primitive assumptions that will guarantee the validity of the additional assumptions in examples.

For a set of vectors a_j, we denote by $(a_j)_{j=1}^4 = (a_1', ..., a_4')'$ a long vector consisting of the four vectors stacked upon each other; we denote by $(a_{jj^*})_{j<j^*}$ the vectors a_{jj^*} stacked together.

Assumption GAUSSIANITY Assume that the following convergence holds:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix} \sqrt{T}\gamma_i \overline{e}_i \\ (\sqrt{T}\gamma_i u_i^{(j)})_{j=1}^4 \\ (T\overline{e}_i u_i^{(j)})_{j=1}^4 \\ (Tu_i^{(j)}u_i^{(j^*)})_{j<j^*} \end{pmatrix} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i \Rightarrow \xi = \begin{pmatrix} \xi_{\gamma e} \\ (\xi_{\gamma j})_{j=1}^4 \\ (\xi_{ej})_{j=1}^4 \\ (\xi_{j,j^*})_{j<j^*} \end{pmatrix},$$

where ξ is a Gaussian vector with mean zero and covariance Σ_ξ.

Assumption COVARIANCE Assume that $\frac{1}{N}\sum_{i=1}^{N}\xi_i\xi_i' \to^p \Sigma_\xi$, where ξ_i and Σ_ξ are defined in Assumption GAUSSIANITY.

The assumptions we maintained in the previous sections are enough to guarantee that $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i$ is $O_p(1)$. Assumption GAUSSIANITY establishes the asymptotic distribution of that quantity, while

Assumption COVARIANCE allows one to construct valid standard errors. Below we provide sufficient conditions for the two new assumptions in the two leading examples discussed before: one where the observed factors are independent from the errors and the example of factor-driven conditional heteroskedasticity.

Lemma 2 *Assume that Assumption ERRORS* holds, and additionally,*

(i) $E\|F_t\|^8 < \infty$; $E\left\|\frac{1}{|T_j|}\sum_{t\in T_j} F_tF_t' - \Sigma_F\right\| \to 0$;

(ii) $\max_i \|\gamma_i\| < C$;

(iii) $\frac{1}{N}\mathrm{tr}(\mathcal{E}_{N,T}^2) \to a_2$ *and* $\frac{1}{N}\gamma'\mathcal{E}_{N,T}\gamma \to \Gamma_\sigma$, *where* Γ_σ *is a full rank matrix;*

(iv) $\frac{1}{N^2}\sum_{i_1=1}^N\sum_{i_2=1}^N\sum_{i_3=1}^N\sum_{i_4=1}^N |Ee_{i_1t}e_{i_2t}e_{i_3t}e_{i_4t}| < C$;

then Assumption GAUSSIANITY holds. If in addition

$$\|\mathcal{E}_{N,T} - \mathrm{dg}(\mathcal{E}_{N,T})\| \to 0 \quad \text{as } N, T \to \infty,$$

then Assumption COVARIANCE holds as well.

Lemma 3 *Assume we have a setting as in Example 1. Assume additionally that conditions (i) and (ii) of Lemma 2 hold and the following is true:*

(i) $E\left[(\|F_t\|^8 + 1)\|w_t\|^8\right] < \infty$;

(ii) $\frac{1}{N}\sum_{i=1}^N \sigma_i^4 \to \mu_4$ *and* $\frac{1}{N}\sum_{i=1}^N \sigma_i^2\gamma_i\gamma_i' \to \Gamma_\sigma$, *where* Γ_σ *is a full rank matrix.*

Then Assumption GAUSSIANITY holds. If in addition $\Gamma_{\pi\gamma} = 0$, *then Assumption COVARIANCE holds as well.*

Assumption GAUSSIANITY is a result of strengthening moment restrictions (condition (i) in both Lemmas), guaranteeing that the asymptotic covariance matrix is well defined and full rank (condition (iii) in Lemma 2 and condition (ii) in Lemma 3) and further restricting cross-sectional dependence (condition (iv) in Lemma 2).

From a theoretical perspective, the derivation of a proper central limit theorem in a factor model setting with a relatively free cross-sectional dependence structure is a major endeavor for two reasons. The first difficulty here is that the quite unrestrictive structure of the cross-sectional dependence of idiosyncratic error terms e_{it} makes ξ_i cross-sectionally dependent, though the correlation between ξ_i and ξ_{i^*} for $i \neq i^*$ converges to zero for large sample sizes. Without imposing further discipline on the structure of dependence, it is hard to obtain a central limit theorem. Secondly, the components ξ_{ej} and ξ_{j,j^*} are quadratic forms in the original errors. Here we use the asymptotic results established for exactly this setting in a separate paper Anatolyev and Mikusheva (2021). There, we exploit time-series conditional independence of errors to obtain a central limit theorem for cross-sectional sums.

Theorem 3 *Assume that the samples $\{r_{it},\ i=1,...,N,\ t=1,...T\}$ and $\{F_t,\ t=1,...,T\}$ come from a data-generating process that satisfies factor pricing model (2) and Assumptions FACTORS, LOADINGS, ERRORS, GAUSSIANITY and COVARIANCE as both N and T increase to infinity. Assume $k_F \geq k_v$. Then*

$$\widehat{\Sigma}_{4S}^{-1/2}(\widehat{\lambda}_{4S}-\lambda) \Rightarrow N(0, I_k).$$

Theorem 3 suggests the use of t- and Wald statistics for the construction of confidence sets and testing hypotheses about values of the risk premia. These inference procedures are standard and can be implemented using standard econometrics software. From a theoretical perspective, however, the asymptotics of the four-split estimator are not fully standard. Notice that the coefficient $\eta_{v,T}$ on the omitted variable μ_i is random, even asymptotically. This implies that the amount of information contained in the sample, which is used to correct for the omitted-variable problem, is random as well, and thus results in the scale of uncertainty depends of the four-split estimator depends on random variables $\eta_{v,T}$. Theorem 3 shows that a properly constructed proxy for the asymptotic variance restores the asymptotic Gaussianity of the multivariate t-statistic. Another important aspect of Theorem 3 is that inferences or construction of a proxy for the variance do not assume knowledge of the number or identity of strong/weak factors. This is a desirable feature, as we do not have a procedure that can credibly differentiate between weak and strong factors.

Theorem 3 is effective for any choice of instruments governed by a choice of matrices $A_1, ..., A_4$ in the four-split algorithm. Intuitively, the informativeness of a proxy for μ_i constructed as a difference in estimated betas is related to Ω_{vF}, the covariance matrix introduced in Assumption FACTORS, which is not easily estimable. The optimal choice of $A_1, ..., A_4$ is an interesting topic for future research and is beyond the scope of the present paper.

Theorem 2 also states that the difference between $\widehat{\lambda}_{4S}$ and $\widetilde{\lambda}$ is of order $\frac{1}{\sqrt{NT}}$. Typically, $\widetilde{\lambda}$ is infeasible. However, if all observed factors are portfolios themselves and are priced by the same model, then we have $\lambda = EF_t$. In such a case the literature suggests the use of an alternative feasible estimator $\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} F_t$, which in this case is equal to $\widetilde{\lambda}$. Thus, in this special case we have two competing estimators for λ and can create a test for model specification. In particular, the statistic compares the difference between $\widehat{\lambda}_{4S}$ and $\widetilde{\lambda}$ to zero. The proof of Theorem 3 shows that $\widehat{\lambda}_{4S}-\widetilde{\lambda}$ converges to zero at the rate $\sqrt{NT}$, and $\widehat{\Sigma}_{IV}$ is a proper proxy for the variance that delivers a χ^2 asymptotic distribution to the corresponding Wald statistic.

5 Empirical Application

5.1 Simulations using artificial data

The goal of this section is to explore the size of potential deficiency of the two-pass procedure and performance of the four-split and competing estimators in a setting close to a real-life application. We calibrate the data-generating process to match the data set of the monthly returns on $N = 100$

Fama-French portfolios, sorted by size and book-to-market and asset returns' relation with the 3 Fama-French factors (market, SmB, HmL). The data are taken from Kenneth French's web-site. We substitute missing values with zeros. The returns are monthly (not annualized), value-weighted, the excess returns are calculated using one-month Treasury bills. The time span is from 01:1972 to 04:2020, resulting in $T = 580$.

First, we run PCA on the panel of excess returns, call the first three principle components G_t with their loadings γ_i, and the fourth main principle component g_t with loadings ϕ_i. We use the normalization $\sum_{t=1}^{T}(G_t', g_t)'(G_t', g_t) = I_4$, making the variance of each factor $1/T$. We compute sample means of the loadings, μ_γ and μ_ϕ, respectively, and their sample variances, V_γ and v_ϕ. We calculate the sample variance σ_ε^2 of the residuals ε_{it} from a regression on the four main PCs. In order to preserve the relation between PCs and Fama-French factors, we run a regression of F_t on a constant and G_t, obtain intercepts $\eta_{0,F}$, slopes η_F and residual variance matrix Σ_{res}. In order to have a proper comparison consistent with our theoretical results, we measure the strength of a given factor as the total variation in the data it produces; that is, its strength is measured as a sum of squared loadings on that factor times the variance of the factor. The strengths of the four main PCs and Fama-French factors are reported in Table 1. These numbers can serve as a reference for the simulation results below.

Table 1: Fractions of variation explained and strengths of principle components and Fama-French factors in Fama-French data set

	1st PC	2nd PC	3rd PC	4th PC
Fraction of variation explained	73.6%	5.9%	3.0%	1.3%
Strength of principle components	2829	229	116	49
	market	SmB	HmL	
Strength of Fama-French factors	2263	482	230	

Simulation design. We simulate the data by following three steps. In the first step, we strive to match the relation of the principle components and Fama-French factors. We simulate $G_t \sim i.i.d.N(0, I_3/T)$ and $\gamma_i \sim i.i.d.N(\mu_\gamma, V_\gamma)$, and then construct the simulated 'observed factors' by $F_t = \eta_{0,F} + \eta_F G_t + w_t$, where $w_t \sim i.i.d.N(0, \Sigma_{res})$. In the second step, we introduce one more factor, part of which represents a missing factor structure in the errors, by simulating $g_t \sim i.i.d.N(0, 1/T)$ and $\phi_i \sim i.i.d.\ \vartheta_\phi \cdot N(\mu_\phi, v_\phi)$. The parameter ϑ_ϕ indexes the strength of the missing factor. Finally, we introduce one more observed factor, which we label as *mom* because we want to imitate the relation of the momentum factor to the PCs. We simulate $mom_t = \eta_{0,mom} + \eta_{mom} G_t + u_t + v_t$, where $\eta_{0,mom}$ and η_{mom} are coefficients from a regression of the momentum factor on the PCs in the data. The simulated error is composed of two parts: $v_t \sim i.i.d.N(0, \varphi\sigma_{mom}^2)$ is uncorrelated with returns or other factors, while $u_t \sim i.i.d.N(0, (1-\varphi)\sigma_{mom}^2)$ is a part of excess returns and is the reason for using *mom* in pricing the assets. Here, σ_{mom}^2 is the residual variance from the regression of *mom* on

PCs. In all simulations, we set $\varphi = 0.001$. We generate loadings on u_t according to $\delta_i = \frac{\alpha\phi_i/\sqrt{T}+\xi_i}{\sqrt{(1-\varphi)\sigma^2_{mom}}}$, where $\xi_i \sim N(0, \sigma^2_\xi)$, so that they are correlated with the loadings on g_t. Now, by increasing α we can increase the correlation, and by increasing σ^2_ξ we can increase the strength of mom.

In the third step, we generate a cross-section of returns and impose a correct pricing model. We generate the demeaned part of excess returns according to $r^*_{it} = G_t\gamma_i + g_t\phi_i + u_t\delta_i + \epsilon_{it}$, where $\epsilon_{it} \sim i.i.d.N(0, \sigma^2_\varepsilon)$ is an idiosyncratic error. The implied true betas are

$$\beta_i = \text{var}\begin{pmatrix} F_t \\ mom_t \end{pmatrix}^{-1} \text{cov}\left(\begin{pmatrix} F_t \\ mom_t \end{pmatrix}, r^*_{it}\right) = \text{var}\begin{pmatrix} F_t \\ mom_t \end{pmatrix}^{-1} \begin{pmatrix} \eta_F\gamma_i/T \\ \eta_{mom}\gamma_i/T + (1-\varphi)\sigma^2_{mom}\delta_i \end{pmatrix}.$$

The correctly priced excess returns are obtained by adding risk premia: $r_{it} = r^*_{it} + \lambda\beta_i$, where λ is the sample means of the Fama-French and momentum factors.

Methods compared. We compare multiple estimators including the two-pass. The four-split estimator uses $A_1 = ... = A_4 = (1, 0, 0, 0)'$. Giglio and Xiu's (2020) three-pass estimator depends on the number p of the principle components selected; we run it for $p = 4$ and $p = 5$. Lettau and Pelger (2020) do not address the question of estimation of risk premia but instead suggest an alternative way to construct factor mimicking portfolios. Unlike the PCA, which extracts factors based on covariances among returns, their method takes into account the expected returns and can pick up a factor that has a high risk premium but is not strongly correlated with returns (i.e., a weak factor). We combined the ideas in the two mentioned papers by implementing a three-pass estimator that uses Lettau and Pelger's (2020) approach in place of PCA to extract factors in the first step.[5] The theoretical statistical properties of this combined procedure are unknown; we computed the standard errors for these estimates in the same way as for the procedure of Giglio and Xiu (2020). In simulations, we used the value of regularization coefficient $\gamma = 20$ and show the results for $K = 4$ and $K = 5$ factors. We also implemented two-split IV estimation without creating a proxy for the missing factor, which would work in the case of no missing factors, but, as expected, it is not competitive in the current setting.[6]

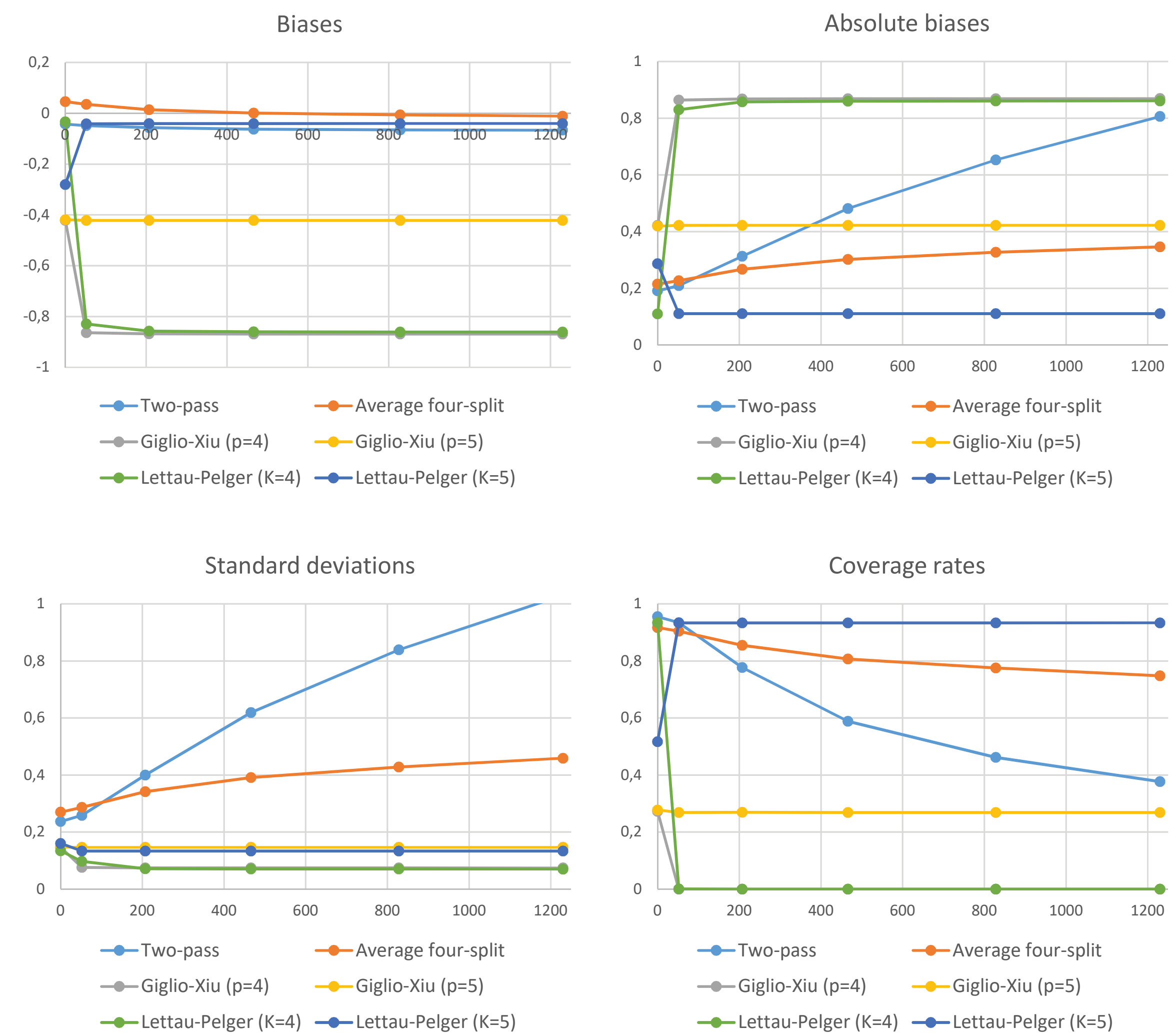
In all simulation experiments, we read off biases, 'absolute' biases, and standard deviations of the estimates for the momentum risk premium, as well as actual 95% coverage rates for the risk premium on the momentum factor. An 'absolute' bias is a characterization of the centrality of a distribution keeping factors fixed. We calculate the bias averaged across $R_{(i)} = 100$ draws of simulated cross sections (such as γ_i, ϕ_i, etc.) keeping factors fixed, then take its absolute value and average across each of $R_{(t)} = 100$ draws of time-series processes (such as G_t, F_t, mom_t, etc.). Thus, the total number of simulations is $R_{(i,t)} = R_{(i)}R_{(t)} = 10,000$.

Simulation results. In the first set of experiments, we explore the effect of a change in the strength of the missing factor. We set $\alpha = 0.1$ and $\sigma^2_\xi = 0.3$, and then vary ϑ_ϕ from 0 to 5, the value

[5] We are grateful to a referee for suggesting this idea.

[6] The results on the two-split estimation are available upon request, we do not report them here in order to remove the clutter.

Figure 1: Simulated statistics for estimates and coverage rates for momentum risk premium

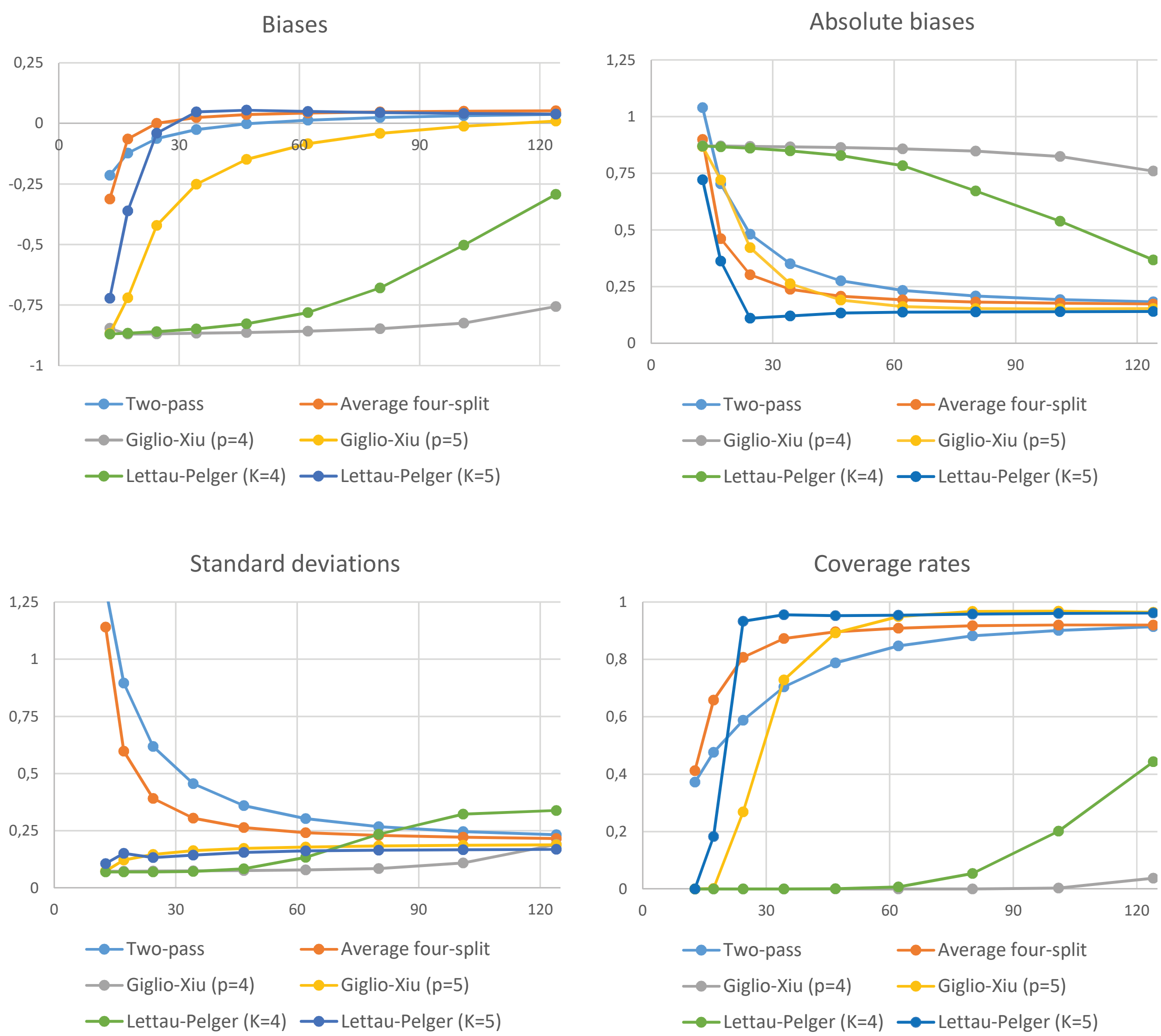


Notes: The strength of factor g_t measured as the total variation explained is on the horizontal axis. The number of simulations of the time series process is $R_{(t)} = 100$. For each time series draw, we simulate $R_{(i)} = 100$ cross-section draws. The total number of simulations is $10,000$.

0 meaning no missing factors, the value 1 corresponding to the strength of the fourth PC. Figure 1 shows the performance measures for different estimators, with the strength of the missing factor (as measured by the total variation explained by the factor, $\frac{1}{T}\sum_{i=1}^{N}\phi_i^2$) on the horizontal axis. One can see that the absolute bias of the two-pass estimator grows fewfold larger than that of the four-split estimator as the missing factor g_t increases in importance. One can see that if the missing factor is of the size of SmB (see Table 1), the two-pass can easily get coverage around 60% instead of the declared 95%, while the four-split largely restores the inferences. The methods of Giglio-Xiu with $p = 4$ and Lettau-Pelger with $k = 4$ factors produce nearly identical results because they extract nearly identical factors (somewhat noisy versions of G_t and g_t) that do not span mom_t. This leads to an extremely biased (though with a low variance) estimator of the risk premium on mom and to

nearly zero coverage of the corresponding confidence sets. It is interesting that these two methods drastically diverge if one extracts 5 factors. Adding the fifth factor to the three-pass with PCA recovers only a small portion of *mom* and still leads to a large bias, but the Lettau-Pelger procedure with 5 factors captures *mom* well due to its non-trivial risk premia. This illustrates the importance of the major assumption underlying the idea of the three-pass procedure – the ability to uncover all priced factors.

Figure 2: Simulated statistics for estimates and coverage rates for momentum risk premium



Notes: The strength of the momentum factor *mom* is on the horizontal axis. The number of simulations of the time series process is $R_{(t)} = 100$. For each time-series draw, we simulate $R_{(i)} = 100$ cross-section draws. The total number of simulations is $10,000$.

In the second set of experiments, we set $\vartheta_\phi = 3$ and $\alpha = 0.1$, and then vary σ_ξ^2 from 0.1 to 0.9 with a step of 0.1, which makes the strength of the missing factor (as measured by $\frac{1}{T}\sum_{i=1}^N \phi_i^2$) fixed at 466, and the strength of the momentum factor (as measured by $\sum_{i=1}^N \beta_{i,4}^2 \frac{1}{T}\sum_{t=1}^T (mom_t - \overline{mom})^2$) increase from 12.7 to 124.0. Figure 2 shows the performance measures for different estimators, with the strength of the momentum factor on the horizontal axis. As *mom* becomes stronger all methods except both three-pass procedures with 4 factors start producing reliable estimate and inferences, especially once the strength of *mom* exceeds 60. Both three-pass procedures with 4 factors fail to uncover the pricing factors correlated with *mom*, although, as we can see, as *mom* become stronger the performance of both methods somewhat improves. The Lettau-Pelger method extracts the factor correlated with *mom* faster than the PCA as it has a non-trivial risk premium. This distinction gets clear when we compare the three-pass procedures with 5 factors extracted by the PCA method and with 5 factors extracted by the Lettau-Pelger procedure. When we compare the four-split with three-pass using the Lettau-Pelger procedure with 5 factors, we notice that the four-split has a higher average variance for all levels of factor strength as probably should be expected from an IV estimator. What is puzzling and surprising of the Lettau-Pelger version of three-pass with 5 factors is that the variance is not reflective of the difficulty of estimation in that it stays nearly constant, which occurs in a stark contrast to the bias which is very high when *mom* is weak. This leads to misleading confidence sets for this method when *mom* has strength lower than 20. The coverage of the four-split also deteriorates for weak *mom*, to a much lesser degree though. In such situations, it may be advisable to use weak IV robust tests at the IV step of four-split estimation.

5.2 Size of the effect in empirical application

In this subsection we run all the estimation procedures described above on two data sets, which we describe further below. We estimate the risk premia on 3 Fama-French factors (market, SmB, HmL) and the momentum factor. For estimation of the long-run variance of observed factors we use the Newey-West estimator with 4 lags (the results are not sensitive to a lag choice). The four-split estimator uses $A_i = (1, 0, 0, 0)'$, the Lettau and Pelger (2020) procedure uses $\gamma = 20$. Since all the observed factors are tradable, we have an alternative (and the most efficient) estimator of the risk premia, $\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} F_t$, the average excess return. This allows us to discuss the quality of different estimates relative to this benchmark. Also, having two valid estimates with different efficiency allows us to test for a correct specification of the linear pricing model. The specification test based on the Wald statistic is equal to the squared difference between the (two-pass or four-split) estimate and the average factor, weighted by an inverse of the difference in covariance matrices. The validity of such a test comes from the proof of Theorem 3. We do not report the specification test for the three-pass estimates as we are not aware of one.

The first data set contains the monthly returns on $N = 100$ Fama-French portfolios sorted by size and book-to-market; they are described in the previous subsection. The time span is from 01:1972 to 04:2020, resulting in $T = 580$. The results are reported in Table 2. The two-pass procedure produces a very high value for the momentum risk premium, and also falsely overstates the accuracy of that estimate to such an extent that the two-pass procedure strongly rejects the linear pricing model with four factors, while the same specification test does not reject the model with only 3

Table 2: Estimates of monthly risk premia on Fama-French factors and momentum factor and tests of specification, using monthly returns on 100 portfolios sorted by size and book-to-market

↓ Procedure	Market	SMB	HML	MOM	Wald	p-value
average excess return →	0.556	0.128	0.263	0.656		
	0.192	0.124	0.145	0.185		
Two-pass	0.612	0.099	0.311	1.905	22.96	0.000
	0.194	0.131	0.152	0.365		
Four-split	0.533	0.151	0.292	0.515	0.81	0.937
	0.198	0.142	0.159	0.804		
Giglio-Xiu ($p = 4$)	0.471	0.206	0.235	−0.155		
	0.237	0.127	0.147	0.105		
Giglio-Xiu ($p = 5$)	0.471	0.207	0.235	−0.155		
	0.243	0.128	0.147	0.106		
Lettau-Pelger ($K = 4$)	0.401	0.253	0.242	0.048		
	0.198	0.124	0.142	0.123		
Lettau-Pelger ($K = 5$)	0.399	0.255	0.242	0.048		
	0.198	0.124	0.143	0.123		

Notes: The sample size is $T = 580$. The standard errors are computed using the Newey-West estimator with 4 lags. In the Giglio and Xiu (2020) three-pass PCA method, p is a maximum number of latent factors to use. In the Lettau and Pelger (2020) RP-PCA procedure, K is a number of factors; the risk premium penalty γ is set to 20.

Fama-French factors. We attribute this to the momentum factor having only weak correlation with returns. The four-split procedure, however, produces an estimate of the momentum risk premium close to the average excess return and accepts the correctness of specification. The four-split estimates have much larger standard errors in comparison to the average excess returns, which is an implicit confirmation of the weakness of the momentum factor. All the variations of the three-pass produce momentum risk premium estimates very far from the benchmark average return on the momentum. This can be explained by the failure of both the PCA and Lettau–Pelger procedure to extract a risk factor correlated with the momentum. The situation here is that the test assets cannot capture all

Table 3: Estimates of monthly risk premia on Fama-French factors and momentum factor and tests of specification, using monthly returns on 25 portfolios sorted by size and book-to-market and 25 portfolios sorted by size and momentum

↓ Procedure	Market	SMB	HML	MOM	Wald	p-value
average excess return →	0.654	0.199	0.331	0.658		
	0.166	0.100	0.119	0.140		
Two-pass	0.673	0.167	0.410	0.740	5.67	0.225
	0.167	0.115	0.136	0.148		
Four-split	0.652	0.237	0.373	0.715	2.73	0.605
	0.169	0.112	0.135	0.151		
Giglio-Xiu ($p = 4$)	0.690	0.307	−0.033	0.427		
	0.379	0.102	0.093	0.172		
Giglio-Xiu ($p = 5$)	0.652	0.202	0.364	0.661		
	0.249	0.094	0.120	0.149		
Lettau-Pelger ($K = 4$)	0.608	0.230	0.218	0.595		
	0.182	0.095	0.092	0.150		
Lettau-Pelger ($K = 5$)	0.592	0.178	0.358	0.646		
	0.174	0.091	0.113	0.140		

Notes: The sample size is $T = 1120$. The standard errors are computed using the Newey-West estimator with 4 lags. In the Giglio and Xiu (2020) three-pass PCA method, p is a maximum number of latent factors to use. In the Lettau and Pelger (2020) RP-PCA procedure, K is a number of factors; the risk premium penalty γ is set to 20.

the pricing information that is necessary to explain the momentum risk premium, but the four-split is still able to uncover some information related to it, though the signal is very weak here as suggested by very large four-split standard errors.

The second data set is intended to be more informative in capturing the momentum. It contains 25 size and book-to-market sorted portfolios and 25 size and momentum sorted portfolios taken from Kenneth French's web-site, hence $N = 50$. The time span is from 01:1927 to 04:2020, leading to $T = 1120$. The results are reported in Table 3. Here all the methods are in agreement. The best results for the three-pass approach is achieved with 5 factors. Interestingly, the four-split exhibits almost no loss of efficiency in comparison to the benchmark and other estimators. Another fact worth pointing out is that a small mistake in estimating a number of factors (like running the three-pass with 4 PCA factors) can yield huge bias while it has almost no reflection on the reported variance.

Acknowledgements

We thank the Editor Serena Ng, the Associate Editor and two anonymous referees for helpful suggestions that have greatly improved the presentation. We appreciate research assistance of David Hughes and Filip Staněk. We also thank audiences of research seminars in MIT-Harvard, Stanford, Duke, UCLA, UC-Davis, USC, Max-Planck Institut, as well as audiences of the 71st Econometric Society European Meeting, the 10th Nordic Econometric Meeting, the 3rd Vienna Workshop on High-Dimensional Time Series in Macroeconomics and Finance, the 4th International Conference on Applied Theory, Macro and Empirical Finance, the 5th International Symposium in Computational Economics and Finance, the 31st European Meeting of Statisticians, and the NES 25th Anniversary Conference. We have benefitted from discussions with Svetlana Bryzgalova, Denis Chetverikov, Tim Christensen, Jinyong Hahn, Roger Moon, and Markus Pelger. Stanislav Anatolyev gratefully acknowledges the Czech Science Foundation support under grants 17-26535S and 20-28055S; Anna Mikusheva gratefully acknowledges the National Science Foundation support under grant 1757199.

References

Anatolyev, S. (2012): "Inference in regression models with many regressors," *Journal of Econometrics*, 170(2), 368-382.

Anatolyev, S. and Mikusheva, A. (2021): "Limit theorems for factor models," *Econometric Theory*, forthcoming. Available at `https://pages.nes.ru/sanatoly/Papers/CLTfm.htm`

Andrews, D.W.K. (2005): "Cross-section regression with common shocks," *Econometrica*, 73(5), 1551-1585.

Angrist, J.D., G.W. Imbens and A. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14, 57-67.

Bai, J. and Ng, S. (2002): “Determining the number of factors in approximate factor models,” *Econometrica*, 70, 191-221.

Bai, J. and S. Ng (2006): “Confidence intervals for diffusion index forecast and inference with factor-augmented regressions,” *Econometrica*, 74, 1133-1155.

Bekker, P.A. (1994): “Alternative approximations to the distributions of instrumental variable estimators,” *Econometrica*, 62, 657-681.

Bryzgalova S. (2016): “Spurious factors in linear asset pricing models,” manuscript, Stanford Graduate School of Business.

Burnside, C. (2016): “Identification and inference in linear stochastic discount factor models with excess returns,” *Journal of Financial Econometrics*, 14(2), 295-330.

Chamberlain, G. and Rothschild, M. (1983): “Arbitrage, factor structure, and mean-variance analysis on large asset markets,” *Econometrica*, 51, 1281-1304.

Cochrane, J. (2001): *Asset Pricing*, Princeton University Press.

Dufour, J.M. and Jasiak, J. (2001): “Finite sample limited information inference methods for structural equations and models with generated regressors,” *International Economic Review*, 42, 815-844.

Fama, E.F. and French, K.R. (1993): “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, 33(1), 3-56.

Fama, E.F. and MacBeth, J. (1973): “Risk, return and equilibrium: Empirical tests,” *Journal of Political Economy*, 81, 607-636.

Feng, G., Giglio, S. and Xiu, D. (2020): “Taming the factor zoo: A test of new factors,” *Journal of Finance*, 75, 1327-1370.

Gagliardini, P., Ossola, E. and Scaillet, O. (2016) “Time-varying risk premium in large cross-sectional equity data sets,” *Econometrica*, 86, 985-1046.

Gagliardini, P., Ossola, E. and Scaillet, O. (2019) “A diagnostic criterion for approximate factor structure,” *Journal of Econometrics*, 212(2), 503-521.

Giglio, S. and Xiu, D. (2020): “Asset pricing with omitted factors,” *Journal of Political Economy*, forthcoming.

Gospodinov, N., Kan, R. and Robotti, C. (2017): “Spurious inference in reduced-rank asset-pricing models,” *Econometrica*, 85(5), 1613-1628.

Hansen, C., Hausman, J. and Newey, W.K. (2008): “Estimation with many instrumental variables,” *Journal of Business & Economics Statistics*, 26, 398-422.

Harvey, C.R., Liu, Y. and Zhu, H. (2016): “... and the cross-section of expected returns,” *Review of Financial Studies*, 29(1), 5-68.

Jagannathan, R. and Wang, Z. (1996): “The conditional CAPM and the cross-section of expected returns. *Journal of Finance*, 51(1), 3-53.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., and Wang, J. (2019): “Empirical tests of asset pricing models with individual assets: Resolving the error-in-variables bias in risk premium estimation,” *Journal of Financial Economics*, 133(2), 273-298.

Jegadeesh, N. and Titman, S. (1993): "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, 48(1), 65-91.

Kan, R. and Zhang, C. (1999): "Two-pass tests of asset pricing models with useless factors," *Journal of Finance*, 54, 203-235.

Kim, S. and Skoulakis, G. (2018): "Ex-post risk premia estimation and asset pricing tests using large cross sections: The regression-calibration approach," *Journal of Econometrics*, 24, 159-188.

Kleibergen, F. (2009): "Tests of risk premia in linear factor models," *Journal of Econometrics*, 149(2), 149-173.

Kleibergen, F., L. Kong, and Z. Zhan (2020): "Identification robust testing of risk premia in finite samples," *Journal of Financial Econometrics*, conditionally accepted.

Kleibergen, F. and Zhan, Z. (2015): "Unexplained factors and their effects on second pass R-squared's," *Journal of Econometrics*, 189, 101-116.

Kozak, S., Nagel, S. and Santosh, S. (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135(2), 271-292.

Lettau, M. and Ludvigson, S. (2001): "Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying," *Journal of Political Economy*, 109(6), 1238-1287.

Lettau, M. and Pelger, M. (2020): " Estimating latent asset-pricing factors," *Journal of Econometrics*, 218(1), 1-242.

Lintner, J. (1965): "Security prices, risk, and maximal gains from diversification," *Journal of Finance*, 20, 587-615.

Lewellen, J., Nagel, S., and Shanken, J. (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, 96, 175-194.

Newey, W.K. and Windmeijer, F. (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77, 687-719.

Onatski, A. (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, 168, 244-258.

Pástor, L. and Stambaugh, R.F. (2003): "Liquidity risk and expected stock returns," *Journal of Political Economy*, 111(3), 642-685.

Raponi, V., Robotti, C., and Zaffaroni, P. (2020): "Testing beta-pricing models using large cross-sections," *Review of Financial Studies*, 33(6), 2796-2842.

Shanken, J. (1992): "On the estimation of beta-pricing models," *Review of Financial Studies*, 5, 1-33.

Sharpe, W.F. (1964): "Capital asset prices: A theory of market equilibrium under conditions of risk," *Journal of Finance*, 19, 425-442.

Staiger, D. and Stock, J.H. (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65(3), 557-586.

Appendix A: proofs

Note that Lemma S1 often referred to is contained in the Supplemental Appendix.

Proof of Theorem 1. Assumption FACTORS guarantees that $\sqrt{T}(\widetilde{\lambda}-\lambda)\Rightarrow N(0,\Omega_F)$. The first-pass (time series) regression yields equation (3), where we use Assumption FACTORS, and $o_p(1)$ appears from the difference between Σ_F and $T^{-1}\sum_{t=1}^T \widetilde{F}_t\widetilde{F}_t'$.

Denote $Q_T = \begin{pmatrix} I_{k_1} & 0_{k_1,k_2} \\ 0_{k_2,k_1} & \sqrt{T}I_{k_2}\end{pmatrix}$. Notice that $Q_T/\sqrt{T}\to\mathcal{I}_{k_2}$. Below we prove the following statement: as $N,T\to\infty$,

$$\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\widehat{\beta}_i'Q_T \Rightarrow (I_{k_\beta};\widetilde{\eta})\Gamma(I_{k_\beta};\widetilde{\eta})' + \mathcal{I}_{k_2}\Sigma_u\mathcal{I}_{k_2}. \tag{6}$$

Indeed,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\widehat{\beta}_i'Q_T &= \frac{1}{N}\sum_{i=1}^N\left(Q_T\beta_i + Q_T\frac{\eta_T}{\sqrt{T}}\mu_i + Q_Tu_i\right)\left(Q_T\beta_i + Q_T\frac{\eta_T}{\sqrt{T}}\mu_i + Q_Tu_i\right)' \\ &= \frac{1}{N}\sum_{i=1}^N\left((I_{k_\beta};\widetilde{\eta}_T)\gamma_i + Q_Tu_i\right)\left((I_{k_\beta};\widetilde{\eta}_T)\gamma_i + Q_Tu_i\right)',\end{aligned} \tag{7}$$

where $\widetilde{\eta}_T = Q_T\eta_T/\sqrt{T}\Rightarrow\mathcal{I}_{k_2}\eta = \widetilde{\eta}$ is $k_F\times k_v$ Gaussian random matrix. Let us show that

$$\frac{T}{N}\sum_{i=1}^N u_iu_i'\to\Sigma_u. \tag{8}$$

Indeed, due to statement (4) of Lemma S1 we have that $\frac{1}{TN}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1,s\neq t}^T\widetilde{F}_t\widetilde{F}_s'e_{it}e_{is} = o_p(1)$. Thus,

$$\begin{aligned}\frac{T}{N}\sum_{i=1}^N u_iu_i' &= \Sigma_F^{-1}\left(\frac{1}{TN}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T\widetilde{F}_t\widetilde{F}_s'e_{it}e_{is}\right)\Sigma_F^{-1} \\ &= \Sigma_F^{-1}\left(\frac{1}{TN}\sum_{i=1}^N\sum_{t=1}^T\widetilde{F}_t\widetilde{F}_t'e_{it}^2\right)\Sigma_F^{-1} + o_p(1)\to^p\Sigma_u,\end{aligned}$$

where the last convergence comes from statement (3) of Lemma S1. Statement (5) of Lemma S1 implies

$$\frac{\sqrt{T}}{N}\sum_{i=1}^N\gamma_iu_i'\to^p 0_{k,k_F}. \tag{9}$$

Combination of equations (7)–(9) and Assumption LOADINGS implies (6).

For the "attenuation bias,"

$$\begin{pmatrix} \sqrt{T}B_1^A \\ B_2^A \end{pmatrix} = Q_T^{-1}\sqrt{T}B^A = -\left(\frac{1}{N}Q_T\sum_{i=1}^N \widehat{\beta}_i\widehat{\beta}_i' Q_T\right)^{-1}\frac{Q_T}{\sqrt{T}}\frac{T}{N}\sum_{i=1}^N u_i u_i'\widetilde{\lambda}.$$

Combining equations (6), (8), $\widetilde{\lambda} \to^p \lambda$ and $Q_T/\sqrt{T} \to \mathcal{I}_{k_2}$, we arrive at

$$\begin{pmatrix} \sqrt{T}B_1^A \\ B_2^A \end{pmatrix} \Rightarrow -\Big((I_{k_\beta};\widetilde{\eta})\Gamma(I_{k_\beta};\widetilde{\eta})' + \mathcal{I}_{k_2}\Sigma_u\mathcal{I}_{k_2}\Big)^{-1}\mathcal{I}_{k_2}\Sigma_u\lambda.$$

For the "omitted variable bias,"

$$\begin{pmatrix} \sqrt{T}B_1^{OV} \\ B_2^{OV} \end{pmatrix} = Q_T^{-1}\sqrt{T}B^{OV} = \left(\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\widehat{\beta}_i' Q_T\right)^{-1}\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\mu_i'\big(\eta_{v,T} - \eta_T'\widetilde{\lambda}\big).$$

Let us consider the following expression:

$$\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\mu_i' = \frac{1}{N}\sum_{i=1}^N \left(Q_T\beta_i + Q_T\frac{\eta_T\mu_i}{\sqrt{T}} + Q_T u_i\right)\mu_i'. \tag{10}$$

By Assumption LOADINGS, $\frac{1}{N}\sum_{i=1}^N Q_T\beta_i\mu_i' \to \Gamma_{\beta\mu}$ and $\frac{1}{N}\sum_{i=1}^N \mu_i\mu_i' \to \Gamma_{\mu\mu}$, while $Q_T\eta_T/\sqrt{T} \Rightarrow \widetilde{\eta}$. The last term in equation (10) is $o_P(1)$ by statement (5) of Lemma S1. Thus,

$$\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\mu_i' \Rightarrow \Gamma_{\beta\mu} + \widetilde{\eta}\Gamma_{\mu\mu}.$$

We also note that $\eta_{v,T} - \eta_T'\widetilde{\lambda} \Rightarrow \eta_v - \eta'\lambda$. This implies validity of the asymptotic statement about B^{OV} contained in Theorem 1.

For the remaining part, by time averaging equation (2) we get $\overline{r}_i = \beta_i'\widetilde{\lambda} + \mu_i'\frac{\eta_{v,T}}{\sqrt{T}} + \overline{e}_i$. Combining the last equation with equation (3), we obtain

$$\overline{r}_i = \widehat{\beta}_i'\widetilde{\lambda} - u_i'\widetilde{\lambda} + \frac{\mu_i'}{\sqrt{T}}\big(\eta_{v,T} - \eta_T'\widetilde{\lambda}\big) + \overline{e}_i.$$

Thus, we arrive at

$$\begin{aligned}\widehat{\lambda}_{TP} - \widetilde{\lambda} - B^A - B^{OV} &= \left(\sum_{i=1}^N \widehat{\beta}_i\widehat{\beta}_i'\right)^{-1}\left(-\sum_{i=1}^N \big(\widehat{\beta}_i - u_i\big)u_i'\widetilde{\lambda} + \sum_{i=1}^N \widehat{\beta}_i\overline{e}_i\right) \\ &= \left(\sum_{i=1}^N \widehat{\beta}_i\widehat{\beta}_i'\right)^{-1}\left(-\sum_{i=1}^N \Big(\beta_i + \frac{\eta_T\mu_i}{\sqrt{T}} + o_p(1)\Big)u_i'\widetilde{\lambda} + \sum_{i=1}^N \widehat{\beta}_i\overline{e}_i\right),\end{aligned}$$

so

$$
\begin{aligned}
&\sqrt{NT}Q_T^{-1}(\widehat{\lambda}_{TP} - \widetilde{\lambda} - B^A - B^{OV}) \\
&\quad = \left(\frac{1}{N}\sum_{i=1}^N Q_T\widehat{\beta}_i\widehat{\beta}_i' Q_T\right)^{-1}\sqrt{\frac{T}{N}}\left(-\sum_{i=1}^N Q_T(\beta_i + \frac{\eta_T\mu_i}{\sqrt{T}})u_i'\widetilde{\lambda} + \sum_{i=1}^N Q_T\widehat{\beta}_i\overline{e}_i\right).
\end{aligned}
$$

Let us prove that the numerator is asymptotically $O_p(1)$:

$$
\begin{aligned}
&\sqrt{\frac{T}{N}}\left(-\sum_{i=1}^N Q_T(\beta_i + \frac{\eta_T\mu_i}{\sqrt{T}})u_i'\widetilde{\lambda} + \sum_{i=1}^N Q_T\widehat{\beta}_i\overline{e}_i\right) \\
&\qquad = (I_{k_\beta}; \widetilde{\eta}_T)\sqrt{\frac{T}{N}}\sum_{i=1}^N \gamma_i(\overline{e}_i - u_i'\widetilde{\lambda}) + \sqrt{\frac{T}{N}}\sum_{i=1}^N Q_T u_i\overline{e}_i + O_p(1). \qquad (11)
\end{aligned}
$$

By statement (5) of Lemma S1, we have $\sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^N \gamma_i\overline{e}_i = O_p(1)$ and $\sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^N \gamma_i u_i' = O_p(1)$, which makes the first summand in equation (11) $O_p(1)$. Consider the second term in equation (11) and recall that $Q_T/\sqrt{T} = O(1)$:

$$
\begin{aligned}
\sqrt{\frac{T}{N}}\sum_{i=1}^N Q_T u_i\overline{e}_i &= \frac{Q_T}{\sqrt{T}}\Sigma_F^{-1}\frac{1}{\sqrt{N}T}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T \widetilde{F}_s e_{is}e_{it} \\
&= \frac{Q_T}{\sqrt{T}}\Sigma_F^{-1}\frac{1}{\sqrt{N}T}\sum_{i=1}^N\sum_{t=1}^T\sum_{s\neq t} \widetilde{F}_s e_{is}e_{it} + \frac{Q_T}{\sqrt{T}}\Sigma_F^{-1}\frac{\sqrt{N}}{T}\sum_{t=1}^T \widetilde{F}_t S_t.
\end{aligned}
$$

The first term is $O_p(1)$ by statement (4) of Lemma S1, while the second term is $O_p(1)$ by Assumption ERRORS(iii). This ends the proof of Theorem 1. □

Proof of Theorem 2. We first discuss the asymptotics of just one IV regression described on step (2), then this argument will be repeated for the other three IV regressions from step (2) of the algorithm. Denote $\tau = \lfloor \frac{T}{4} \rfloor = |T_j|$.

The time-series regression on a sub-sample j gives us that

$$
\widehat{\beta}_i^{(j)} = \left(\beta_i + u_i^{(j)} + \frac{\eta_{j,T}\mu_i}{\sqrt{\tau}}\right)(1 + o_p(1)),
$$

where $\eta_{j,T} = \frac{1}{\sqrt{\tau}}\sum_{t\in T_j}\Sigma_F^{-1}\widetilde{F}_t^{(j)}v_t' \Rightarrow \eta_j$, η_j is random $k_F \times k_v$ matrix with the distribution $\mathrm{vec}(\eta_j) \sim N(0_{k_F k_v,1}, \Omega_{vF})$, and the $o_p(1)$ term is related to the difference between Σ_F and $\frac{1}{\tau}\sum_{t\in T_j}\widetilde{F}_t^{(j)}\widetilde{F}_t^{(j)\prime}$.

On step (2) we run an IV regression of $y_i = \frac{1}{T}\sum_{t=1}^T r_{it}$ on the regressor

$$
x_i^{(1)} = \begin{pmatrix} \widehat{\beta}_i^{(1)} \\ A_1\big(\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_i^{(1)} \\ A_1\frac{\eta_{1,T}-\eta_{2,T}}{\sqrt{\tau}}\mu_i + A_1\big(u_i^{(1)} - u_i^{(2)}\big) \end{pmatrix},
$$

with the instruments

$$z_i^{(1)} = \begin{pmatrix} \widehat{\beta}_i^{(3)} \\ \widehat{\beta}_i^{(3)} - \widehat{\beta}_i^{(4)} \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_i^{(3)} \\ \frac{\eta_{3,T}-\eta_{4,T}}{\sqrt{\tau}}\mu_i + (u_i^{(3)} - u_i^{(4)}) \end{pmatrix}.$$

The main estimation equation can be written in the following way:

$$\begin{aligned} y_i &= \frac{1}{T}\sum_{t\in T} r_{it} = \widetilde{\lambda}'\beta_i + \frac{\eta_{v,T}'}{\sqrt{T}}\mu_i + \overline{e}_i = \widetilde{\lambda}'\widehat{\beta}_i^{(1)} + \left(\frac{\eta_{v,T}'}{\sqrt{T}} - \widetilde{\lambda}'\frac{\eta_{1,T}}{\sqrt{\tau}}\right)\mu_i + \overline{e}_i - \widetilde{\lambda}'u_i^{(1)} \\ &= \widetilde{\lambda}'\widehat{\beta}_i^{(1)} + a_{1,T}A_1\big(\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}\big) + \overline{e}_i - \widetilde{\lambda}'u_i^{(1)} - a_{1,T}A_1\big(u_i^{(1)} - u_i^{(2)}\big). \end{aligned}$$

Thus, we can write it as follows:

$$y_i = (\widetilde{\lambda}', a_{1,T})x_i^{(1)} + \epsilon_i^{(1)}. \tag{12}$$

Here we use the following notation:

$$a_{1,T} = \left(\frac{\eta_{v,T}'}{\sqrt{T}} - \widetilde{\lambda}'\frac{\eta_{1,T}}{\sqrt{\tau}}\right)\left(A_1\frac{\eta_{1,T}-\eta_{2,T}}{\sqrt{\tau}}\right)^{-1} \Rightarrow \left(\frac{\eta_v'}{2} - \eta_1\right)\big(A_1(\eta_1 - \eta_2)\big)^{-1},$$

and $\epsilon_i^{(1)} = \overline{e}_i - \widetilde{\lambda}'u_i^{(1)} - a_{1,T}A_1\big(u_i^{(1)} - u_i^{(2)}\big)$. Notice that $a_{1,T}$ is a random $1 \times k_v$ matrix that is well defined with probability approaching 1 (as $\eta_{1,T}$ and $\eta_{2,T}$ weakly converge to two independent random Gaussian matrices), and $a_{1,T}$ is asymptotically of order $O_p(1)$.

The estimator computed on the step (2) of the four-split algorithm is

$$\widehat{\lambda}^{(1)} = \big(I_{k_F}, 0_{k_F,k_v}\big)\left(X^{(1)\prime}Z^{(1)}\big(Z^{(1)\prime}Z^{(1)}\big)^{-1}Z^{(1)\prime}X^{(1)}\right)^{-1}X^{(1)\prime}Z^{(1)}\big(Z^{(1)\prime}Z^{(1)}\big)^{-1}Z^{(1)\prime}Y.$$

Using equation (12) we obtain:

$$\widehat{\lambda}^{(1)} - \widetilde{\lambda} = \big(I_{k_F}, 0_{k_F,k_v}\big)\left(X^{(1)\prime}P_{Z^{(1)}}X^{(1)}\right)^{-1}X^{(1)\prime}P_{Z^{(1)}}\epsilon^{(1)}, \tag{13}$$

where P_Z is a projection matrix onto Z. Let us introduce two normalizing matrices:

$$Q_x = \begin{pmatrix} Q_T & 0_{k_F,k_v} \\ 0_{k_v,k_F} & \sqrt{T}I_{k_v} \end{pmatrix}, \qquad Q_z = \begin{pmatrix} Q_T & 0_{k_F,k_F} \\ 0_{k_F,k_F} & \sqrt{T}I_{k_F} \end{pmatrix}.$$

The dimensionality of Q_x is $k \times k$, where $k = k_F + k_v$ is a number of regressors in the second stage regression, while the dimensionality of Q_z is $2k_F \times 2k_F$, where $2k_F$ is the number of instruments. The matrix Q_T was defined in the proof of Theorem 1. Now,

$$Q_x x_i^{(1)} = \left(\widetilde{A}_{1,T}\gamma_i + \begin{pmatrix} Q_T & 0_{k_F,k_F} \\ \sqrt{T}A_1 & -\sqrt{T}A_1 \end{pmatrix}\begin{pmatrix} u_i^{(1)} \\ u_i^{(2)} \end{pmatrix}\right),$$

where

$$\widetilde{A}_{1,T} = \begin{pmatrix} I_{k_F} & Q_T \frac{\eta_{1,T}}{\sqrt{\tau}} \\ 0_{k_v,k_F} & 2A_1(\eta_{1,T} - \eta_{2,T}) \end{pmatrix} \Rightarrow \begin{pmatrix} I_{k_F} & 2\mathcal{I}_{k_2}\eta_1 \\ 0_{k_v,k_F} & 2A_1(\eta_1 - \eta_2) \end{pmatrix} = \widetilde{A}_1,$$

$$\frac{1}{\sqrt{T}} \begin{pmatrix} Q_T & 0_{k_F,k_F} \\ \sqrt{T}A_1 & -\sqrt{T}A_1 \end{pmatrix} \to \begin{pmatrix} \mathcal{I}_{k_2} & 0_{k_F,k_F} \\ A_1 & -A_1 \end{pmatrix}.$$

Here $\mathcal{I}_{k_2}$ is a $k_F \times k_F$ matrix which was introduced in Theorem 1. We also have

$$Q_z z_i^{(1)} = \left(A_{1,T}^* \gamma_i + \begin{pmatrix} Q_T/\sqrt{T} & 0_{k_F,k_F} \\ I_{k_F} & -I_{k_F} \end{pmatrix} \sqrt{T} \begin{pmatrix} u_i^{(3)} \\ u_i^{(4)} \end{pmatrix} \right),$$

where

$$A_{1,T}^* = \begin{pmatrix} I_{k_F} & Q_T \frac{\eta_{3,T}}{\sqrt{\tau}} \\ 0_{k_F,k_F} & 2(\eta_{3,T} - \eta_{4,T}) \end{pmatrix} \Rightarrow \begin{pmatrix} I_{k_F} & 2\mathcal{I}_{k_2}\eta_3 \\ 0_{k_F,k_F} & 2(\eta_3 - \eta_4) \end{pmatrix} = A_1^*,$$

$$\begin{pmatrix} Q_T/\sqrt{T} & 0_{k_F,k_F} \\ I_{k_F} & -I_{k_F} \end{pmatrix} \to \begin{pmatrix} \mathcal{I}_{k_2} & 0_{k_F,k_F} \\ I_{k_F} & -I_{k_F} \end{pmatrix}.$$

Statements (1) and (5) of Lemma S1 imply that

$$\frac{T}{\sqrt{N}} \sum_{i=1}^{N} u_i^{(j)} u_i^{(j^*)\prime} = O_p(1) \quad \text{for} \quad j \neq j^*, \tag{14}$$

$$\sqrt{\frac{T}{N}} \sum_{i=1}^{N} (\gamma_i', 1)' u_i^{(j^*)\prime} = O_p(1). \tag{15}$$

This together with Assumption LOADINGS gives us that

$$\frac{1}{N} \sum_{i=1}^{N} Q_x x_i^{(1)} z_i^{(1)\prime} Q_z \Rightarrow \widetilde{A}_1 \Gamma A_1^{*\prime}. \tag{16}$$

By Assumption LOADINGS, Γ is full rank, while $\widetilde{A}_1$ and $A_1^{*\prime}$ are full rank with probability 1. Statements (3) and (4) of Lemma S1 imply that

$$\frac{\tau}{N} \sum_{i=1}^{N} u_i^{(j)} u_i^{(j)\prime} \to^p \Sigma_u. \tag{17}$$

Thus, we obtain

$$\frac{1}{N}\sum_{i=1}^{N} Q_z z_i^{(1)} z_i^{(1)\prime} Q_z \Rightarrow A_1^* \Gamma A_1^{*\prime} + 4\begin{pmatrix} \mathcal{I}_{k_2} & 0_{k_F,k_F} \\ I_{k_F} & -I_{k_F} \end{pmatrix}\begin{pmatrix} \Sigma_u & 0_{k_F,k_F} \\ 0_{k_F,k_F} & \Sigma_u \end{pmatrix}\begin{pmatrix} \mathcal{I}_{k_2} & I_{k_F} \\ 0_{k_F,k_F} & -I_{k_F} \end{pmatrix}$$
$$= A_1^* \Gamma A_1^{*\prime} + 4\begin{pmatrix} \mathcal{I}_{k_2}\Sigma_u \mathcal{I}_{k_2} & \mathcal{I}_{k_2}\Sigma_u \\ \Sigma_u \mathcal{I}_{k_2} & 2\Sigma_u \end{pmatrix}. \quad (18)$$

Let us now show that

$$\sqrt{\frac{T}{N}}\sum_{i=1}^{N} Q_z z_i^{(1)} \epsilon_i^{(1)} = O_p(1). \quad (19)$$

We have $\epsilon_i^{(1)} = \overline{e}_i - \widetilde{\lambda}' u_i^{(1)} - a_{1,T} A_1 (u_i^{(1)} - u_i^{(2)})$. The sum in (19) contains summands of the form $\sqrt{\frac{T}{N}} \sum_{i=1}^N \gamma_i(\overline{e}_i, u_i^{(j)})$, $\frac{T}{\sqrt{N}} \sum_{i=1}^N \overline{e}_i u_i^{(j)}$ and $\frac{T}{\sqrt{N}} \sum_{i=1}^N u_i^{(j^*)\prime} u_i^{(j)}$. All three types of summands are $O_p(1)$ due to statements (5), (2) and (1) of Lemma S1, correspondingly. Putting equations (16) and (18) together, we obtain

$$NQ_x^{-1}\Theta_{N,T,1}Q_z^{-1} = \left(\frac{Q_x X^{(1)\prime} Z^{(1)} Q_z}{N} \left(\frac{Q_z Z^{(1)\prime} Z^{(1)} Q_z}{N} \right)^{-1} \frac{Q_z Z^{(1)\prime} X^{(1)} Q_x}{N} \right)^{-1}$$
$$\cdot \frac{Q_x X^{(1)\prime} Z^{(1)} Q_z}{N} \left(\frac{Q_z Z^{(1)\prime} Z^{(1)} Q_z}{N} \right)^{-1} = O_p(1).$$

Putting everything together, we have:

$$\sqrt{NT} Q_T^{-1} (\widehat{\lambda}^{(1)} - \widetilde{\lambda}) = (I_{k_F}, 0_{k_F,k_v}) N Q_x^{-1} \Theta_{N,T,1} Q_z^{-1} \sqrt{\frac{T}{N}} \sum_{i=1}^{N} Q_z z_i^{(1)} \epsilon_i^{(1)} = O_p(1).$$

Because $\sqrt{NT} Q_T^{-1} = \begin{pmatrix} \sqrt{NT} I_{k_1} & 0_{k_1,k_2} \\ 0_{k_2,k_1} & \sqrt{N} I_{k_2} \end{pmatrix}$, we obtain different rates of estimation of the risk premia λ_1 and λ_2 on the strong and weak observed factors. We have $\sqrt{NT}\big(\widehat{\lambda}_1^{(1)} - \widetilde{\lambda}_1\big) = O_p(1)$, while $\sqrt{N}\big(\widehat{\lambda}_2^{(1)} - \widetilde{\lambda}_2\big) = O_p(1)$. Thus,

$$\sqrt{T}\big(\widehat{\lambda}_1^{(1)} - \lambda_1\big) = \sqrt{T}\big(\widetilde{\lambda}_1 - \lambda_1\big) + \sqrt{T}\big(\widehat{\lambda}_1^{(1)} - \widetilde{\lambda}_1\big) = \sqrt{T}\big(\widetilde{\lambda}_1 - \lambda_1\big) + O_p\big(1/\sqrt{N}\big) \Rightarrow N(0, \Omega_F).$$

As for the estimator of the risk premia on the weak factors,

$$\widehat{\lambda}_2^{(1)} - \lambda_2 = \big(\widetilde{\lambda}_2 - \lambda_2\big) + \big(\widehat{\lambda}_2^{(1)} - \widetilde{\lambda}_2\big) = O_p\big(1/\sqrt{T}\big) + O_p\big(1/\sqrt{N}\big) = O_p\big(1/\sqrt{\min\{N,T\}}\big).$$

We have proved the statement of Theorem 2 for an estimator obtained on step (2) of the algorithm, but the same line of reasoning applies to $\widehat{\lambda}^{(2)}, \widehat{\lambda}^{(3)}, \widehat{\lambda}^{(4)}$ and their average. This finishes the proof of Theorem 2. □

Proof of Theorem 3. Following the steps of the proof of Theorem 2 we get the following two statements:

$$\frac{1}{N}\sum_{i=1}^{N} Q_x x_i^{(j)} z_i^{(j)\prime} Q_z \Rightarrow \widetilde{A}_j \mathbf{\Gamma} A_j^{*\prime}, \tag{20}$$

$$\frac{1}{N}\sum_{i=1}^{N} Q_z z_i^{(j)} z_i^{(j)\prime} Q_z \Rightarrow A_j^{*} \mathbf{\Gamma} A_j^{*\prime} + 4\begin{pmatrix} \mathcal{I}_{k_2}\Sigma_u\mathcal{I}_{k_2} & \mathcal{I}_{k_2}\Sigma_u \\ \Sigma_u\mathcal{I}_{k_2} & 2\Sigma_u \end{pmatrix}, \tag{21}$$

where A_j^* and $\widetilde{A}_j$ are random matrices that are deterministic functions of random vectors $(\eta_1, ..., \eta_4)$. Indeed, let us adopt the following notation. Let $j_1, ..., j_4$ be the circular indexes used for computing $\widehat{\lambda}^{(j)}$. In particular, the estimate $\widehat{\lambda}^{(j)}$ is computed from the IV regression with the regressors $x_i^{(j)} = \big(\widehat{\beta}_i^{(j_1)\prime}, (\widehat{\beta}_i^{(j_1)} - \widehat{\beta}_i^{(j_2)})' A_j'\big)'$ and the instruments $z_i^{(j)} = \big(\widehat{\beta}_i^{(j_3)\prime}, (\widehat{\beta}_i^{(j_3)} - \widehat{\beta}_i^{(j_4)})'\big)'$. Then, similarly to the proof of Theorem 2, we obtain:

$$A_j^* = \begin{pmatrix} I_{k_F} & 2\mathcal{I}_{k_2}\eta_{j_3} \\ 0_{k_F,k_F} & 2(\eta_{j_3} - \eta_{j_4}) \end{pmatrix}, \qquad \widetilde{A}_j = \begin{pmatrix} I_{k_F} & 2\mathcal{I}_{k_2}\eta_{j_1} \\ 0_{k_v,k_F} & 2A_j(\eta_{j_1} - \eta_{j_2}) \end{pmatrix}.$$

So,

$$NQ_x^{-1}\left(X^{(j)\prime}Z^{(j)}\big(Z^{(j)\prime}Z^{(j)}\big)^{-1}Z^{(j)\prime}X^{(j)}\right)^{-1} X^{(j)\prime}Z^{(j)}\big(Z^{(j)\prime}Z^{(j)}\big)^{-1}Q_z^{-1} \Rightarrow \Theta_j.$$

The limit Θ_j in the last expression is a known deterministic function of random vectors $(\eta_1, ..., \eta_4)$, which can be explicitly written in terms of A_j^* and $\widetilde{A}_j$.

We have the following expression for the estimates obtained on steps (2) and (3) of the four-split algorithm:

$$\sqrt{NT}Q_T^{-1}\big(\widehat{\lambda}^{(j)} - \widetilde{\lambda}\big) = \big(I_{k_F}, 0_{k_F,k_v}\big)NQ_x^{-1}\Theta_{N,T,j}Q_z^{-1}\sqrt{\frac{T}{N}}\sum_{i=1}^{N} Q_z z_i^{(j)}\epsilon_i^{(j)},$$

where $\epsilon_i^{(j)} = \overline{e}_i - \widetilde{\lambda}' u_i^{(j_1)} - a_{j,T}A_j\big(u_i^{(j_1)} - u_i^{(j_2)}\big)$, and

$$Q_z z_i^{(j)} = A_{j,T}^*\gamma_i + \begin{pmatrix} Q_T/\sqrt{T} & 0_{k_F,k_F} \\ I_{k_F} & -I_{k_F} \end{pmatrix}\sqrt{T}\begin{pmatrix} u_i^{(j_3)} \\ u_i^{(j_4)} \end{pmatrix}.$$

The following term can be rewritten in terms of ξ_i from Assumption GAUSSIANITY:

$$\sqrt{\frac{T}{N}}\sum_{i=1}^{N} Q_z z_i^{(j)}\epsilon_i^{(j)} = A_{j,T}^{*}\left(\sqrt{\frac{T}{N}}\sum_{i=1}^{N}\gamma_i\begin{pmatrix}\overline{e}_i\\ u_i^{(j_1)}\\ u_i^{(j_2)}\end{pmatrix}'\right)\begin{pmatrix}1\\ -\widetilde{\lambda}-A_j'a_{j,T}'\\ A_j'a_{j,T}'\end{pmatrix}$$
$$+\begin{pmatrix}Q_T/\sqrt{T} & 0_{k_F,k_F}\\ I_{k_F} & -I_{k_F}\end{pmatrix}\left(\frac{T}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix}u_i^{(j_3)}\\ u_i^{(j_4)}\end{pmatrix}\begin{pmatrix}\overline{e}_i\\ u_i^{(j_1)}\\ u_i^{(j_2)}\end{pmatrix}'\right)\begin{pmatrix}1\\ -\widetilde{\lambda}-A_j'a_{j,T}'\\ A_j'a_{j,T}'\end{pmatrix}$$
$$=\mathcal{A}_{j,T}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i,$$

where $\mathcal{A}_{j,T}$ is a $k_z \times k_\xi$ matrix which is a deterministic function of $A_{j,T}^{*}, A_j, a_{j,T}, \widetilde{\lambda}$. The exact expression for $\mathcal{A}_{j,T}$ is obvious though too complicated to write down. We have discussed before the convergence of all terms separately, which implies that $\mathcal{A}_{j,T} \Rightarrow \mathcal{A}_j$, where the limit is a deterministic function of $(\eta_1, ..., \eta_4)$.

Given Assumption GAUSSIANITY, we have $\sqrt{\frac{T}{N}}\sum_{i=1}^{N} Q_z z_i^{(j)}\epsilon_i^{(j)} \Rightarrow \mathcal{A}_j\xi$. Following step (4) of the four-split algorithm, we can put all pieces together:

$$\sqrt{NT}Q^{-1}\big(\widehat{\lambda}_{4S}-\widetilde{\lambda}\big) \Rightarrow \big(I_{k_F}, 0_{k_F,k_v}\big)\left(\frac{1}{4}\sum_{j=1}^{4}\Theta_j\mathcal{A}_j\right)\xi. \tag{22}$$

As we can see, the four-split estimator is asymptotically mixed Gaussian; that is, the limit distribution conditionally on $\eta_1, ..., \eta_4$ (which is independent of ξ due to Assumption ERRORS) is Gaussian with mean zero and the variance depending on $\eta_1, ..., \eta_4$.

Denote $\widehat{\Sigma}_{IV} = \frac{1}{N}R'G^{-1}\widehat{\Sigma}_0 G^{-1}R$. We show below that $\widehat{\Sigma}_{IV}$ has the following asymptotic distribution:

$$NTQ_T^{-1}\widehat{\Sigma}_{IV}Q_T^{-1} \Rightarrow \big(I_{k_F}, 0_{k_F,k_v}\big)\left(\frac{1}{4}\sum_{j=1}^{4}\Theta_j\mathcal{A}_j\right)\Sigma_\xi\left(\frac{1}{4}\sum_{j=1}^{4}\Theta_j\mathcal{A}_j\right)'\big(I_{k_F}, 0_{k_F,k_v}\big)'. \tag{23}$$

Statement (23) implies the statement of Theorem 3. Indeed, equations (22) and (23) imply that $\widehat{\Sigma}_{IV}^{-1/2}\big(\widehat{\lambda}_{4S}-\widetilde{\lambda}\big) \Rightarrow N(0, I_k)$, where the limiting Gaussian vector is independent of the limiting Gaussian vector in the following convergence:

$$\sqrt{T}\Omega_F^{-1/2}\big(\widetilde{\lambda}-\lambda\big) \Rightarrow N(0, I_k).$$

The expression $\widehat{\Sigma}_{4S}^{-1/2}\big(\widehat{\lambda}_{4S}-\lambda\big)$ is the weighted sum of the expressions staying on the left-hand-side of the last two convergence with weights asymptotically independent from both limiting $N(0, I_k)$. This leads to the validity of the statement of Theorem 3.

To prove the validity of statement (23) we notice that $\sqrt{T}Q_z z_i^{(j)}\epsilon_i^{(j)} = \mathcal{A}_{j,T}\xi_i$. Thus,

$$\frac{T}{N}\sum_{i=1}^{N}\begin{pmatrix} Q_z\widetilde{z}_i^{(1)}\epsilon_i^{(1)} \\ \dots \\ Q_z\widetilde{z}_i^{(4)}\epsilon_i^{(4)}\end{pmatrix}\begin{pmatrix} Q_z\widetilde{z}_i^{(1)}\epsilon_i^{(1)} \\ \dots \\ Q_z\widetilde{z}_i^{(4)}\epsilon_i^{(4)}\end{pmatrix}' = \begin{pmatrix}\mathcal{A}_{1,T} \\ \dots \\ \mathcal{A}_{4,T}\end{pmatrix}\frac{1}{N}\sum_{i=1}^{N}\xi_i\xi_i'\begin{pmatrix}\mathcal{A}_{1,T} \\ \dots \\ \mathcal{A}_{4,T}\end{pmatrix}'$$

$$\Rightarrow \begin{pmatrix}\mathcal{A}_{1} \\ \dots \\ \mathcal{A}_{4}\end{pmatrix}\Sigma_\xi\begin{pmatrix}\mathcal{A}_{1} \\ \dots \\ \mathcal{A}_{4}\end{pmatrix}'. \tag{24}$$

Let us consider an infeasible variance estimator $\widetilde{\Sigma}_{IV}$ which is constructed in the same way as $\widehat{\Sigma}_{IV}$ but uses $\epsilon_i^{(j)}$ in place of $\widehat{\epsilon}_i^{(j)}$. That is, denote

$$\widetilde{\Sigma}_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\widetilde{z}_i^{(1)}\epsilon_i^{(1)} \\ \dots \\ \widetilde{z}_i^{(4)}\epsilon_i^{(4)}\end{pmatrix}\begin{pmatrix}\widetilde{z}_i^{(1)}\epsilon_i^{(1)} \\ \dots \\ \widetilde{z}_i^{(4)}\epsilon_i^{(4)}\end{pmatrix}',$$

and consider $\widetilde{\Sigma}_{IV} = \frac{1}{N}R'G^{-1}\widetilde{\Sigma}_0 G^{-1}R$. By putting together (20), (21) and (24) we obtain

$$NTQ_T^{-1}\widetilde{\Sigma}_{IV}Q_T^{-1} \Rightarrow \left(I_{k_F}, 0_{k_F,k_v}\right)\left(\frac{1}{4}\sum_{j=1}^{4}\Theta_j\mathcal{A}_j\right)\Sigma_\xi\left(\frac{1}{4}\sum_{j=1}^{4}\Theta_j\mathcal{A}_j\right)'\left(I_{k_F}, 0_{k_F,k_v}\right)'.$$

The only thing left to show is that the difference between $\widehat{\Sigma}_{IV}$ and $\widetilde{\Sigma}_{IV}$ is asymptotically negligible. In particular, we will show for any j and j^*,

$$\frac{T}{N}\sum_{i=1}^{N}Q_z z_i^{(j)}z_i^{(j^*)\prime}Q_z\left(\epsilon_i^{(j)}\epsilon_i^{(j^*)} - \widehat{\epsilon}_i^{(j)}\widehat{\epsilon}_i^{(j^*)}\right) \to^p 0, \tag{25}$$

where $\widehat{\epsilon}_i^{(j)}$ are the residuals from the $(j)^{th}$ IV regression. Indeed, this last statement implies that $\widehat{\Sigma}_{IV} = \widetilde{\Sigma}_{IV}\big(1 + o_p(1)\big)$, and usage of residuals in place of true errors does not have an asymptotic effect on estimation of variance.

In order to prove (25) we write down an equation analogous to equation (12):

$$y_i = (\widetilde{\lambda}', a_{j,T})x_i^{(j)} + \epsilon_i^{(j)} = \theta_j' x_i^{(j)} + \epsilon_i^{(j)}.$$

From the proof of Theorem 2 we have that $\sqrt{NT}Q_x^{-1}\big(\widehat{\theta}_j - \theta_j\big) = O_p(1)$, where $\widehat{\theta}_j$ is the IV estimator obtained on Steps (2) for $j = 1$ or on Step (3) for $j = 2, 3$ or 4. The residuals for this regression are

$$\widehat{\epsilon}_i^{(j)} = y_i - \widehat{\theta}_j' x_i^{(j)} = \epsilon_i^{(j)} - \big(\widehat{\theta}_j - \theta_j\big)' x_i^{(j)} = \epsilon_i^{(j)} - \big(Q_x^{-1}\big(\widehat{\theta}_j - \theta_j\big)\big)' Q_x x_i^{(j)}.$$

The left hand expression of (25) is equal to

$$\frac{T}{N}\sum_{i=1}^{N} Q_z z_i^{(j)} z_i^{(j^*)\prime} Q_z \left(\epsilon_i^{(j)}\big(\widehat{\theta}_{j^*}-\theta_{j^*}\big)' x_i^{(j^*)} + \epsilon_i^{(j^*)}\big(\widehat{\theta}_j-\theta_j\big)' x_i^{(j)} - \big(\widehat{\theta}_{j^*}-\theta_{j^*}\big)' x_i^{(j^*)}\big(\widehat{\theta}_j-\theta_j\big)' x_i^{(j)}\right).$$

This expression contains three sums. We can show that each of them is asymptotically negligible. For example, consider the first of the three sums:

$$\begin{aligned}
&\frac{1}{N^{3/2}}\sum_{i=1}^{N}\big(\sqrt{T} Q_z z_i^{(j)}\epsilon_i^{(j)}\big)\big(Q_z z_i^{(j^*)}\big)'\big(Q_x x_i^{(j^*)}\big)'\sqrt{NT} Q_x^{-1}\big(\widehat{\theta}_{j^*}-\theta_{j^*}\big)\\
&\quad = \frac{1}{N^{3/2}}\sum_{i=1}^{N}\mathcal{A}_{j,T}\xi_i\big(Q_z z_i^{(j^*)}\big)'\big(Q_x x_i^{(j^*)}\big)'\sqrt{NT} Q_x^{-1}\big(\widehat{\theta}_{j^*}-\theta_{j^*}\big).
\end{aligned}$$

Note that $\sqrt{NT}Q_x^{-1}\big(\widehat{\theta}_{j^*}-\theta_{j^*}\big) = O_p(1)$. As before, $Q_z z_i^{(j)} = O_p(1)\gamma_i + O_p(1)\sqrt{T}\big(u_i^{(j_3)}, u_i^{(j_4)}\big)'$, while $Q_x x_i^{(j)} = O_p(1)\gamma_i + O_p(1)\sqrt{T}\big(u_i^{(j_1)}, u_i^{(j_2)}\big)$, where all the mentioned $O_p(1)$ terms are not indexed by i. Thus,

$$\frac{1}{N^{3/2}}\sum_{i=1}^{N}\mathcal{A}_{j,T}\xi_i\big(Q_z z_i^{(j^*)}\big)'\big(Q_x x_i^{(j^*)}\big)' = O_p(1)\frac{1}{N^{3/2}}\sum_{i=1}^{N}\xi_i\xi_i' + O_p(1)\frac{1}{N^{3/2}}\sum_{i=1}^{N}\xi_i\otimes(\gamma_i\gamma_i').$$

By Assumption GAUSSIANITY, $\frac{1}{N^{3/2}}\sum_{i=1}^{N}\xi_i\xi_i' \to^p 0$ and thus

$$\frac{1}{N^{3/2}}\left\|\sum_{i=1}^{N}\xi_i\otimes(\gamma_i\gamma_i')\right\| \leq \frac{1}{N^{3/2}}\sqrt{\sum_{i=1}^{N}\|\xi_i\|^2}\sqrt{\sum_{i=1}^{N}\|\gamma_i\|^4} \to^p 0.$$

This gives the asymptotic negligibility of the first sum; the negligibility of the other two sums is proved in a similar manner. This ends the proof of Theorem 3. □