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

Stanislav Anatolyev[1] Anna Mikusheva[2]

[1]CERGE-EI and NES

[2]MIT

December 2017

Linear factor-pricing models

- Factor-pricing model:

$$Er_{it} = \lambda'\beta_i, \text{ where } \beta_i = \mathbf{var}(F_t)^{-1}\mathbf{cov}(F_t, r_{it})$$

  r_{it} is excess return to portfolio i at period t, F_t are risk factors, β_i are risk exposures, λ are risk premia.

- Classical estimation approach is the two-pass procedure (Fama and MacBeth, 1973) with standard error correction (Shanken, 1992)
  1. Estimate β_i for each portfolio from time-series regression;
  2. Estimate λ from cross-sectional regression of average returns on estimated betas.
- Quality control:
  - Is price of risk non-zero? Test: $H_0 : \lambda \neq 0$;
  - Do these risks price market? Specification test $H_0 : Er_{it} = \lambda'\beta_i$;
  - How much does risk exposure explain a variation in average returns? Second-pass R^2.

Linear factor-pricing models

- First and most known: CAPM (Sharpe 1964, Linner 1965)
- The second most well-known is Fama-French (1993): includes market portfolio, size factor 'SMB' (small-minus-big) and book-to-market factor 'HML'(high-minus-low).
- Some models have factors based on market behavior: example-momentum factor 'MOM'(Jegadeesh and Titman, 1993);
- Some have macroeconomic factors: example- consumption-to-wealth ratio 'cay' (Lettau and Ludvigson, 2001)
- Harvey, Liu and Zhu (2016) list hundreds of papers proposing, justifying and estimating various linear factor-pricing models.

Problem 1: weak identification?

- If some of the observed factors are only weakly correlated with returns, then the second-pass parameters may be weakly identified.
  - Kan and Zhang (1999): useless factors lead to spurious inference
  - Kleibergen and Zhan (2015): weak factors may arise from poor measurement of true factors
  - Kleibergen (2009): weak factors distort consistency and asymptotic normality of risk-premia estimates.

Problem 2: missing factors?

- Empirical fact found in Kleibergen and Zhan (2015): many well-known linear factor-pricing models have very strong remaining factor structure present in the residuals.
- Example: for all Lettau and Ludvigson (2001) specifications first three principle components of residuals explain 82% - 96% of remaining cross-sectional variation.
- One found exception to this rule: Fama and French.

Observation in our paper: Large T and large N?

- Traditionally (and in all mentioned papers) the asymptotic results are derived under assumption:

$$N \text{ is fixed}, \ T \to \infty$$

- However, the most often used datasets are:
  - Jagannathan and Wang (1996): $N = 100$, $T = 330$;
  - Fama-French: $N = 25$, $T = 141$;
  - Gagliardini, Ossola and Scaillet (2016): $N = 44$ and $N = 9936$, $T = 546$.
- N and T are comparable in size
- More adequate asymptotic approximations may result from both $N \to \infty$ and $T \to \infty$

Our setup includes simultaneously

- **Weak observed factors:** Some observed factors are only weakly correlated: we model corresponding risk exposure coefficients β_i as being of order $O(1/\sqrt{T})$. Thus, first-stage estimation error is of the same order of magnitude as the coefficients themselves
- **Missing factors:** There is a strong factor structure present in error terms
- **Large-N-large-T asymptotics:** Many assets-long time span:

$$N, T \to \infty$$

Findings of our paper

- We prove that the classical two-pass procedure fails in our setting: inconsistent estimates of the premia on weak factors, invalid inferences and significant finite-sample bias for estimate of risk premia on strong observed factor
- We propose new procedures that provide consistent estimators for risk premia and guarantee asymptotically gaussian inferences.

Findings of our paper

- We develop an estimation procedure for risk premia in an environment with many assets, weak included factors and strong excluded factors with the following features:
    - it yields consistent estimates when the traditional two-pass procedure fails;
    - it yields consistent estimates without knowledge of which factors are strong and which are weak;
    - it does not lose efficiency if the traditional two-pass procedure works;
    - it is a procedure of the ‘press button’ type: easy-to-implement, uses standard estimation techniques.

Outline

Setup

- We observe excess returns on assets or portfolios $\{r_{it}, i = 1, ..., N, t = 1, ..., T\}$ and $k_F \times 1$ risk factors $\{F_t, t = 1, ..., T\}$ that follow the correctly-specified linear factor-pricing model:

$$Er_{it} = \lambda'\beta_i, \text{ where } \beta_i = \mathbf{var}(F_t)^{-1}\mathbf{cov}(F_t, r_{it})$$

- This is equivalent to assuming that

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

where the random error terms ε_{it} have mean zero and are uncorrelated with F_t. We treat λ and β_i as non-random, while $r_{it}, F_t, \varepsilon_{it}$ are random.

Setup: weak observed factors

- We will divide factors $F_t = (F'_{t,1}, F'_{t,2})'$ and exposures $\beta_i = (\beta'_{i,1}, \beta'_{i,2})'$ into "strong" and "weak":
  - $\beta_{i,2} = \frac{b_i}{\sqrt{T}}$, where we make the same assumptions about size of $\beta_{i,1}$ and size of b_i (they are $O(1)$).
  - Estimation error for each β_i is of order $O_p(1/\sqrt{T})$, similar to size of $\beta_{i,2}$
  - In setting with N-fixed and $T \to \infty$, this corresponds to weak identification.
- We do not assume that econometrician knows which factors are weak or the number of weak factors (our results hold for more general assumptions, that some linear combination of factors is weak).

Setup: missing factors

- Model:

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

- We assume that error terms are not auto-correlated (efficient market hypothesis) but have non-trivial cross-sectional dependence - they have unobserved factor structure:

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

where

  - v_t are unobserved random variables; have mean zero and unit variance (normalization); uncorrelated with e_{it};
  - μ_i - unknown constant loadings of size $O(1)$.
  - e_{it} are weakly cross-sectionally correlated.

Outline

Asymptotics of the two-pass procedure

- If all observed factors are strong: $\Rightarrow \sqrt{T}(\widehat{\lambda}_{TP} - \lambda) \Rightarrow N(0, V)$.
- If some observed factors are weak, but no missing factors in errors: $\Rightarrow$ "errors-in-variables" bias:
  - $\widehat{\lambda}_{TP,1}$ is consistent and Gaussian, but biased (inferences are not valid),
  - $\widehat{\lambda}_{TP,2}$ is inconsistent
- If some observed factors are weak, and some missing factors in errors: $\Rightarrow$ "errors-in-variables" + "omitted variable":
  - $\widehat{\lambda}_{TP,1}$ is consistent, but biased and non-standard distribution,
  - $\widehat{\lambda}_{TP,2}$ is inconsistent

Why two-pass fails? No missing factors case

- Assume some observed factors are weak, but no factor structure in errors
$$r_{it} = \lambda'\beta_i + (F_t - EF_t)'\beta_i + e_{it},$$
- e_{it} are weakly dependent
- First-pass estimates:

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

where $u_i = \frac{1}{T}\sum_{t=1}^{T} \Sigma_F^{-1}\widetilde{F}_t e_{it}$ are 'asymptotically uncorrelated' for different i and unrelated to β_i

Why two-pass fails? No missing factors case

- Ideal regression: if one regresses $\bar{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_{it}$ on β_i, then will have consistent estimate of λ
- But we have instead only estimates and $u_i = O(1/\sqrt{T})$

$$\begin{pmatrix} \widehat{\beta}_{i,1} \\ \widehat{\beta}_{i,2} \end{pmatrix} = \begin{pmatrix} \beta_{i,1} \\ \beta_{i,2} \end{pmatrix} + \begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} = \begin{pmatrix} \beta_{i,1}(1+o(1)) \\ \beta_{i,2} + u_{i,2} \end{pmatrix}$$

- Mistake in $\beta_{i,2}$ is of the same order of magnitude as coefficient itself. It behaves like classical measurement error!
- Regression of $\bar{r}_i$ on $\widehat{\beta}_i$ has an attenuation bias!

No missing factors case: Solution

Idea:

- Split sample in two $T_1 \sqcup T_2 = \{1, ..., T\}$
- Estimate β_i twice:

$$\widehat{\beta}_i^{(j)} = \left(\sum_{t \in T_j} \widetilde{F}_t \widetilde{F}_t' \right)^{-1} \sum_{t \in T_j} \widetilde{F}_t r_{it} = (\beta_i + u_i^{(j)})(1 + o_p(1)), \quad j = 1, 2$$

- Estimation mistakes $u_i^{(1)}$ and $u_i^{(2)}$ are (asymptotically) uncorrelated
- Use $\widehat{\beta}_i^{(1)}$ as a regressor and $\widehat{\beta}_i^{(2)}$ as instrument (or vice versa, or both and average final estimates)
- Idea of sample-splitting (and its extreme version: leave-one-out or jackknife) has been used in many-weak-IV model (Hansen, Hausman and Newey, 2008)

Factors in errors. Why two-pass fails?

- Model with factor structure in errors:

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

  v_t is unobserved and μ_i are unknown, e_{it} are weakly cross-correlated.

- First step

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

  where

$$\eta_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Sigma_F^{-1} \tilde{F}_t v_t'$$

  is coming from unobserved factor structure

Factors in errors. Why two-pass fails?

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

- Now the estimation error $\frac{\eta_T \mu_i}{\sqrt{T}} + u_i$ is NOT classical measurement error:
    - both terms $\frac{\eta_T \mu_i}{\sqrt{T}}$ and u_i are stochastically of order $O_p(\frac{1}{\sqrt{T}})$
    - estimation errors are cross-correlated (for different i) due to term $\frac{\eta_T \mu_i}{\sqrt{T}}$
    - estimation error may be 'correlated' with regressor if 'sample correlation' between β_i and μ_i is non-zero

Factors in errors. Why two-pass fails?

- Model with factor structure in errors:

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

- Ideal regression:

$$y_i = \sqrt{T}\overline{r}_i = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} r_{it} = \widetilde{\lambda}'\left(\sqrt{T}\beta_i\right) + \eta_v'\mu_i + \varepsilon_i,$$

- If there is μ_i but you know β_i only- we have omitted variable, it will cause omitted variable bias if 'sample correlation' between β_i and μ_i is non-zero.

Factors in errors. Why two-pass fails?

Summary:

- if there is no factor structure in errors - we have classical error-in-variables problem and associated attenuation bias
- If we have factor structure in errors we additionally have:
  - non-classical error-in-variable (mistakes in regressor $\widehat{\beta}_{i,2}$ are cross-correlated and 'correlated' with β_i)
  - even if we know β_i there is omitted variable bias in the 'ideal' regression if 'sample correlation' between β_i and μ_i is non-zero.

Outline

Our proposed solution: Idea

- We reconsider sample-splitting.
- We have an estimate of β_i for each sub-sample

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

where

$$\eta_j = \frac{1}{\sqrt{|T_j|}} \sum_{t \in T_j} \Sigma_F^{-1} \tilde{F}_t v_t' \Rightarrow N(\mathbf{0}, \Omega_{Fv}).$$

- η_j are independent for different j and independent from errors $u_i^{(j)}$.

Our proposed solution: Idea

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

- We can construct proxy for μ_i (!!!)

$$\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)} = \left(\frac{\eta_1}{\sqrt{|T_1|}} - \frac{\eta_2}{\sqrt{|T_2|}}\right)\mu_i + (u_i^{(1)} - u_i^{(2)})$$

- If $|T_j| = T/4$, then 'random' coefficient $\left(\frac{\eta_1}{\sqrt{|T_1|}} - \frac{\eta_2}{\sqrt{|T_2|}}\right) = O(\frac{1}{\sqrt{T}})$ and error $(u_i^{(1)} - u_i^{(2)}) = O(\frac{1}{\sqrt{T}})$
- Proxy $\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}$ mis-measures μ_i, but measurement error is classical: not cross-correlated and not correlated with regressors.

Our proposed solution: Idea

- Split sample into 4 equal sub-samples.
- Estimate $\widehat{\beta}_i^{(j)}$ for $j = 1, ..., 4$.
- Run IV regression of $\overline{r}_i$ on regressors $\widehat{\beta}_i^{(1)}$ and proxy based on $\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(2)}$ with instruments $\widehat{\beta}_i^{(3)}$ and $\widehat{\beta}_i^{(3)} - \widehat{\beta}_i^{(4)}$.
- For efficiency considerations you may repeat this 4 times circulating indices 1-4.
- Average estimates you obtain for λ.
- We also provide formula for how to calculate covariance matrix for our estimate.

Our proposed solution

- The exact asymptotic distribution of $\widehat{\lambda}_{4S}$ is not Gaussian but rather *mixed* Gaussian. The estimated variance matrix is asymptotically random though non-degenerate with probability 1.
- This is due to the fact that the coefficient on proxy for μ_i is random. It leads to information contained in second stage IV being random, though NOT weak with probability 1.
- Our 4-split estimator:
  - it yields consistent estimates when the traditional two-pass procedure fails;
  - it yields consistent estimates without knowledge of which factors are strong and which are weak;
  - it does not lose efficiency if the traditional two-pass procedure works;
  - it is a procedure of the ‘push-button’ type: easy-to-implement, uses standard estimation techniques.

Outline

Empirical application (Fama–French portfolios)

no.	specification	5 main principal components in residuals				
1	Market, SMB, HML	0.29	0.14	0.11	0.07	0.04
2	Market, HML	0.62	0.10	0.05	0.03	0.03
3	Market, HML, cay	0.62	0.10	0.05	0.03	0.03

Empirical application (Fama–French portfolios)

no.	specification	5 main principal components in residuals				
1	Market, SMB, HML	0.29	0.14	0.11	0.07	0.04
2	Market, HML	0.62	0.10	0.05	0.03	0.03
3	Market, HML, cay	0.62	0.10	0.05	0.03	0.03

no.	risk factor	Market	SMB	HML	cay
1	conventional two-pass	2.70 0.61	0.69 0.48	1.96 0.58	
	average four-split	2.80 0.62	0.46 0.47	1.29 0.84	
3	conventional two-pass	2.55 0.61		1.92 0.62	0.027 0.019
	average four-split	2.06 0.63		2.44 0.68	−0.009 0.005

Empirical application (industry portfolios)

specification	5 main principal components in residuals				
Market, SMB, HML, MOM	0.14	0.12	0.08	0.06	0.04

Empirical application (industry portfolios)

specification	5 main principal components in residuals				
Market, SMB, HML, MOM	0.14	0.12	0.08	0.06	0.04

risk factor	Market	SMB	HML	MOM
conventional two-pass	1.05 (0.20)	−0.27 (0.19)	−0.00 (0.15)	1.05 (0.35)
average four-split	1.15 (0.21)	−1.10 (0.24)	0.03 (0.18)	0.03 (0.40)

Conclusion

What we have done here:

- Showed that conventional two-pass procedure gives unreliable estimates of risk premia in empirically-relevant situations
- Proposed alternative "press buttons" procedure robust to weak factors and strong missing factors, based on split-sample IV
- Alternative procedure yields consistent and asymptotically normal estimates under many-asset, weak-factor asymptotics