A diagnostic criterion for approximate factor structure

Patrick Gagliardini[a], Elisa Ossola[b] and Olivier Scaillet[c]

[a]Università della Svizzera Italiana (USI Lugano) and Swiss Finance Institute,
[b]European Commission, Joint Research Centre,
[c]University of Genève and Swiss Finance Institute.

February, 2017

Goal of the paper

- Define a simple diagnostic criterion: (i) for **approximate factor structure** in **large cross-sectional equity datasets**; (ii) to determine the **number of omitted common factors**.
- *Main idea:* If the set of observable factors is correctly specified, the errors are only weakly cross-sectionally correlated.
- Relevant because the no-arbitrage restrictions from APT do not hold when factors are omitted.

Building blocks of the paper

1. A new simple diagnostic criterion

- Conditional linear factor model in a **large economy**, with an **approximate factor structure** for excess returns.
  Gagliardini-Ossola-Scaillet (GOS, 2016).
- **Large unbalanced panel** of returns.
- Cross-section and time-series dimensions: $\mathbf{n}, \mathbf{T} \to \infty$ s.t. either $T/n = o(1)$ or $T/n \to c$, where $c > 0$.
- **Observable** factors.
- Link with criteria for unobservable factors.

  Large balanced panels: Connor-Korajczyk (1993), Stock-Watson (2002);

  Procedure to estimate the number of factors: Bai-Ng (2002); Onatski (2010), Ahn-Horenstein (2013), Caner-Han (2014);

  Procedure for hypotheses on the number of factors: Onatski (2009);

2. Empirical analysis results with CRSP **individual stock returns**

- Use of individual stocks with monthly and quarterly returns: $n >> T$.
- Several linear multi-factor models built by using a large number of empirical factors.
- **Monthly data**:
  - Time invariant and time-varying specifications of the financial factor models.
  - We conclude for no omitted factors in the errors for the **time-invariant models** with at least **four financial factors**.
  - We conclude for no omitted factor structure in the errors for **a scaled three factor Fama-French specification**.
- **Quarterly data**:
  - We cannot select macroeconomic models without **the market factor**.

Outline of the presentation

- Introduction ✓
- Model setting
  - Conditional factor model
  - Rival models
- Diagnostic criterion
- Determining the number of omitted factors
- Empirical results
- Conclusions

Model setting: Conditional linear factor model

Excess returns generation

The excess return $R_t(\gamma)$ of asset $\gamma \in [0,1]$ at date $t = 1, 2, ...,$ satisfies

$$R_t(\gamma) = \beta_t(\gamma)' x_t + \varepsilon_t(\gamma), \tag{1}$$

where:

- $x_t = (1, f_t')'$ and f_t is the $K \times 1$ random vector of observable factors;
- $\beta_t(\gamma) = \left(a_t(\gamma), b_t(\gamma)'\right)'$ contains time-varying coefficients;
- $\varepsilon_t(\gamma)$ is a random vector of error terms s.t. $E[\varepsilon_t(\gamma)|\mathcal{F}_{t-1}] = 0$ and $Cov[\varepsilon_t(\gamma), f_t|\mathcal{F}_{t-1}] = 0$ for any $\gamma \in [0,1]$.

(Hansen-Richard (1987))

- *Approximate factor structure:* (Chamberlain-Rothschild (1983)) **nondiagonal** conditional error var-cov matrix $\Sigma_{\varepsilon,t,n} = [Cov[\varepsilon_t(\gamma_i), \varepsilon_t(\gamma_j)|\mathcal{F}_{t-1}]]_{i,j=1,\ldots,n}$ with bounded largest eigenvalue if model is correct.
- *No asymptotic arbitrage opportunities:* there are no portfolios that approximate arbitrage opportunities when the number of assets increases.
- *Asset pricing restriction:* $a_t(\gamma) = b_t(\gamma)' \nu_t$ holds a.s. in probability (GOS (2016)).
  $\Leftrightarrow E[R_t(\gamma)|\mathcal{F}_{t-1}] = b_t(\gamma)' \lambda_t$, where $\lambda_t = \nu_t + E[f_t|\mathcal{F}_{t-1}]$ is the vector of risk premia.
- *Large economy with a continuum of assets:*
  robustness of factor structures to asset **repackaging** (Al-Najjar (1995, 1998, 1999), GOS (2016)).
- *Unbalanced nature of the panel:*
  $I_t(\gamma)$ admits value 1 if the return of asset γ is observable at date t, and 0 otherwise (Connor-Korajczyk (1987)).

Functional specification of time-varying coefficients

Information set $\mathcal{F}_{t-1}$ contains lagged observations of:

- $Z_t \in \mathbb{R}^p$, vector of common instruments:
  - the constant and observable factors f_t,
  - additional observable variables Z_t^*.
- $Z_t(\gamma) \in \mathbb{R}^q$, vector of asset-specific instruments:
  - firm characteristics,
  - stocks returns.

Factor loadings: $b_t(\gamma) = B(\gamma) Z_{t-1} + C(\gamma) Z_{t-1}(\gamma)$, where $B(\gamma) \in \mathbb{R}^{K \times p}$ and $C(\gamma) \in \mathbb{R}^{K \times q}$, for any $\gamma \in [0,1]$ and $t = 1, 2, ...$;

Risk premia: $\lambda_t = \Lambda Z_{t-1}$, where $\Lambda \in \mathbb{R}^{K \times p}$, for any t;

Factors: $E[f_t | \mathcal{F}_{t-1}] = F Z_{t-1}$, where $F \in \mathbb{R}^{K \times p}$, for any t.

The sampling scheme: (Andrews (2005))
A sample of n assets is obtained by drawing i.i.d. indices γ_i according to a probability distribution G on $[0,1]$.

$\Rightarrow$ **cross-sectional limits exist and are invariant to reordering of assets.**

$\Rightarrow$ sample of n assets and T observations of excess returns
$R_{i,t} = R_t(\gamma_i)$, $I_{i,t} = I_t(\gamma_i)$, $\varepsilon_{i,t} = \varepsilon_t(\gamma_i)$, $Z_{i,t} = Z_t(\gamma)$
for $i = 1, ..., n$ and $t = 1, ..., T$.

$\Rightarrow$ **random coefficient panel** model with $\beta_{i,t} = \beta_t(\gamma_i)$.

- The **conditional factor model** (1), for the sample observations, becomes

$$R_{i,t} = x'_{i,t}\beta_i + \varepsilon_{i,t}, \tag{2}$$

where

  - regressor $x_{i,t}$ involves cross-terms of instruments Z_{t-1}, $Z_{i,t-1}$ and f_t;
  - time-invariant parameters $\beta_i = \left(\beta'_{1,i}, \beta'_{2,i}\right)'$ are (unconditional) transformations of matrices B_i, C_i, Λ and F.

- In matrix notation, for any asset i, we have

$$R_i = X_i\beta_i + \varepsilon_i,$$

where R_i and ε_i are $T \times 1$ vectors.

Rival models

$\mathcal{M}_1$: the linear regression model (2), where the errors $(\varepsilon_{i,t})$ are weakly cross-sectionally correlated (approximate factor structure);
$\mathcal{M}_2$: the linear regression model (2), where the errors $(\varepsilon_{i,t})$ satisfy a factor structure:

$$\varepsilon_{i,t} = \theta_i' h_t + u_{i,t}, \tag{3}$$

where vector h_t includes m unobservable common factors.

In matrix notation, for any asset i, under $\mathcal{M}_2$, we have

$$\varepsilon_i = H\theta_i + u_i,$$

where H is the $T \times m$ matrix of unobservable factor values, and u_i is a $T \times 1$ vector.

Assumption 1: Presence of some common factors in the errors

Under model $\mathcal{M}_2$,

(i) $\underset{T\to\infty}{\text{plim}} \frac{1}{T}\sum_t h_t h_t' = \Sigma_h$, where Σ_h is a positive definite matrix;

(ii) $\mu_1\left(\frac{1}{n}\sum_i \theta_i\theta_i'\right) \geq C$, with probability approaching 1, for a constant $C > 0$,

where $\mu_1(\cdot)$ denotes the largest eigenvalue of a symmetric matrix.

- Assumption 1(i) is a standard condition on the latent factor (equivalent to the Assumption A in Bai-Ng (2002));
- Assumption 1(ii) requires that at least one factor in the error terms is strong (equivalent to the Assumption B in Bai-Ng (2002)).

Assumption 2: Granger non causality assumption

The best linear prediction of the unobservable factor $EL\left(h_t \middle| \{x_{i,t},\ i=1,2,...\}\right)$ is independent of $\{\tilde{x}_{i,t},\ i=1,2,...\}$ where the regressor $x_{i,t} = \left(x_t', \tilde{x}_{i,t}'\right)'$ is decomposed in the common component $x_t = \left(vech\left[X_t\right]', f_t' \otimes Z_{t-1}'\right)$ and the stock-specific component $\tilde{x}_{i,t} = \left(Z_{t-1}' \otimes Z_{i,t-1}', f_t' \otimes Z_{i,t-1}'\right)'$.

- Assumption 2 implies orthogonality between latent factors and observable factors for all stocks, i.e. $E\left[x_{i,t} h_t'\right] = 0,\ \forall i$.
- If $x_{i,t} = x_t$, we get $E\left[x_t h_t'\right] = 0$ by a transformation of the latent factors: *identification restriction*.
- Assumption 2 is mantained under model $\mathcal{M}_2$.

Assumption 3:

The cross-sectional dimension n and time series dimension T are such that $n = O(T^{\bar{\gamma}})$, $\bar{\gamma} > 0$, and $T = O(n^{\gamma})$, $\gamma \in (0,1]$.

Diagnostic criterion

The diagnostic criterion is

$$\xi = \mu_1 \left(\frac{1}{nT} \sum_i \mathbf{1}_i^{\chi} \bar{\varepsilon}_i \bar{\varepsilon}_i' \right) - g\left(n, T\right),$$

where

- the vector $\bar{\varepsilon}_i \in \mathbb{R}^T$ contains $\bar{\varepsilon}_{i,t} = I_{i,t} \hat{\varepsilon}_{i,t}$ with $\hat{\varepsilon}_{i,t} = R_{i,t} - x_{i,t}' \hat{\beta}_i$, and $\hat{\beta}_i$ are estimated by OLS regression on (2) as in GOS (2016);
- the penalty $g\left(n, T\right)$ is such that $g(n, T) \to 0$ and $C_{n,T}^2 g(n, T) \to \infty$, when $n, T \to \infty$, for $C_{n,T}^2 = \min\{n, T\}$.

Proposition 1: Model selection rule

Under Assumptions 1 and 2, when $n, T \to \infty$,
(a) we select $\mathcal{M}_1$ if $\xi < 0$,
(b) we select $\mathcal{M}_2$ if $\xi > 0$,
since $Pr(\mathcal{M}_1 | \xi < 0) \to 1$ and $Pr(\mathcal{M}_2 | \xi > 0) \to 1$.

- In the balanced case: $\xi = SS_0 - SS_1 - g(n, T)$, where
  - SS_0 is the sum of squared residuals,
  - $SS_1 = \min_{H \in \mathbb{R}^T,\ \Theta \in \mathbb{R}^n} \frac{1}{nT} \sum_i \sum_t (\hat{\varepsilon}_{i,t} - \theta_i h_t)^2$, subject to the normalization constraint $\frac{H'H}{T} = 1$.
- Penalized criteria for zero- and one-factor model in Bai-Ng (2002):
$$\xi = PC(0) - PC(1),$$
where $PC(0) = SS_0$, and $PC(1) = SS_1 + g(n, T)$.

The diagnostic criterion based on a *logarithmic transform*:

$$\begin{aligned}\check{\xi} &= \ln\left(\frac{1}{nT}\sum_i\sum_t \mathbf{1}_i^{\chi}\bar{\varepsilon}_{i,t}^2\right) \\ &\quad - \ln\left(\frac{1}{nT}\sum_i\sum_t \mathbf{1}_i^{\chi}\bar{\varepsilon}_{i,t}^2 - \mu_1\left(\frac{1}{nT}\sum_i \mathbf{1}_i^{\chi}\bar{\varepsilon}_i\bar{\varepsilon}_i'\right)\right) - g(n,T).\end{aligned}$$

- In the balanced case: $\check{\xi} = \ln(SS_0/SS_1) - g(n,T)$.
- Information criteria for zero- and one-factor model in Bai-Ng (2002):
$$\check{\xi} = IC(0) - IC(1).$$
- The selection rule is the same as in Proposition 1 with $\check{\xi}$ substituted for ξ.

Do we have one, two, or more omitted factors?

Rival models

$\mathcal{M}_1(k)$: the linear regression model (2), where the errors $(\varepsilon_{i,t})$ satisfy a factor structure with exactly k unobservable factors;
$\mathcal{M}_2(k)$: the linear regression model (2), where the errors $(\varepsilon_{i,t})$ satisfy a factor structure with at least $k+1$ unobservable factors.

Assumption 3: Identification of unobservable factors in the errors

Under model $\mathcal{M}_2(k)$, we have $\mu_{k+1}\left(\frac{1}{nT}\sum_i \theta_i\theta_i'\right) \geq C$, with probability approaching 1, for a constant $C > 0$, where $\mu_{k+1}(\cdot)$ denotes the $(k+1)$-largest eigenvalues of a symmetric matrix.

- Assumption 3 requires that there are at least $k+1$ strong factors under $\mathcal{M}_2(k)$.

Diagnostic criterion

The diagnostic criterion is

$$\xi(k) = \mu_{k+1}\left(\frac{1}{nT}\sum_i \mathbf{1}_i^{\chi}\bar{\varepsilon}_i\bar{\varepsilon}_i'\right) - g(n,T).$$

Proposition 2: Model selection rule

Under Assumptions 1(i), 2 and 3, when $n, T \to \infty$:
(a) we select $\mathcal{M}_1(k)$ if $\xi(k) < 0$, (b) we select $\mathcal{M}_2(k)$ if $\xi(k) > 0$, since $Pr(\mathcal{M}_1(k)|\xi(k) < 0) \to 1$ and $Pr(\mathcal{M}_2(k)|\xi(k) > 0) \to 1$.

The number of omitted factors is
$\hat{k} = \min\{k = 0, 1, ..., T-1 : \xi(k) < 0\}$.

Data description

Base assets:

- 10,442 stocks with monthly and quarterly returns from Jan1968 to Dec2011 after merging CRSP and Compustat databases.

Linear factor specifications:

- involving financial factors: CAPM and Fama-French models, among others;
- involving financial and macro factors: CCAPM and Epstein-Zin model, among others.

Conditional specifications:
(i) common variables: $Z_{t-1} = (1, divY_{t-1})'$;
(ii) common variables: $Z_{t-1} = (1, divY_{t-1})'$ and firm characteristics $Z_{i,t-1} = bm_{i,t-1}$.

Implementation: the rescaled diagnostic criterion

$$\xi(k) = \frac{\mu_{k+1}\left(\frac{1}{nT}\sum_i \mathbf{1}_i^{\chi}\bar{\varepsilon}_i\bar{\varepsilon}_i'\right)}{\hat{\sigma}^2} - g(n,T)$$

measures the contribution (in %) of the $(k+1)$th eigenvalue to the variance $\hat{\sigma}^2 = \frac{1}{nT}\sum_i\sum_t \mathbf{1}_i^{\chi}\bar{\varepsilon}_{i,t}^2 = \sum_{j=1}^{T}\mu_j\left(\frac{1}{nT}\sum_i \mathbf{1}_i^{\chi}\bar{\varepsilon}_i\bar{\varepsilon}_i'\right)$.

- Rescaled eigenvalues are **easier to interpret**!
- In practice, we standardize each time series of residuals $\bar{\varepsilon}_i$ to have unit variance, i.e. $\bar{\bar{\varepsilon}}_{i,t} = \frac{\bar{\varepsilon}_{i,t}}{\sqrt{\frac{1}{T}\sum_t \bar{\varepsilon}_{i,t}^2}}$ (see Bai-Ng (2002)) and we compute

$$\xi(k) = \mu_{k+1}\left(\frac{1}{n^{\chi}T}\sum_i \mathbf{1}_i^{\chi}\bar{\bar{\varepsilon}}_i\bar{\bar{\varepsilon}}_i'\right) - g(n^{\chi},T),$$

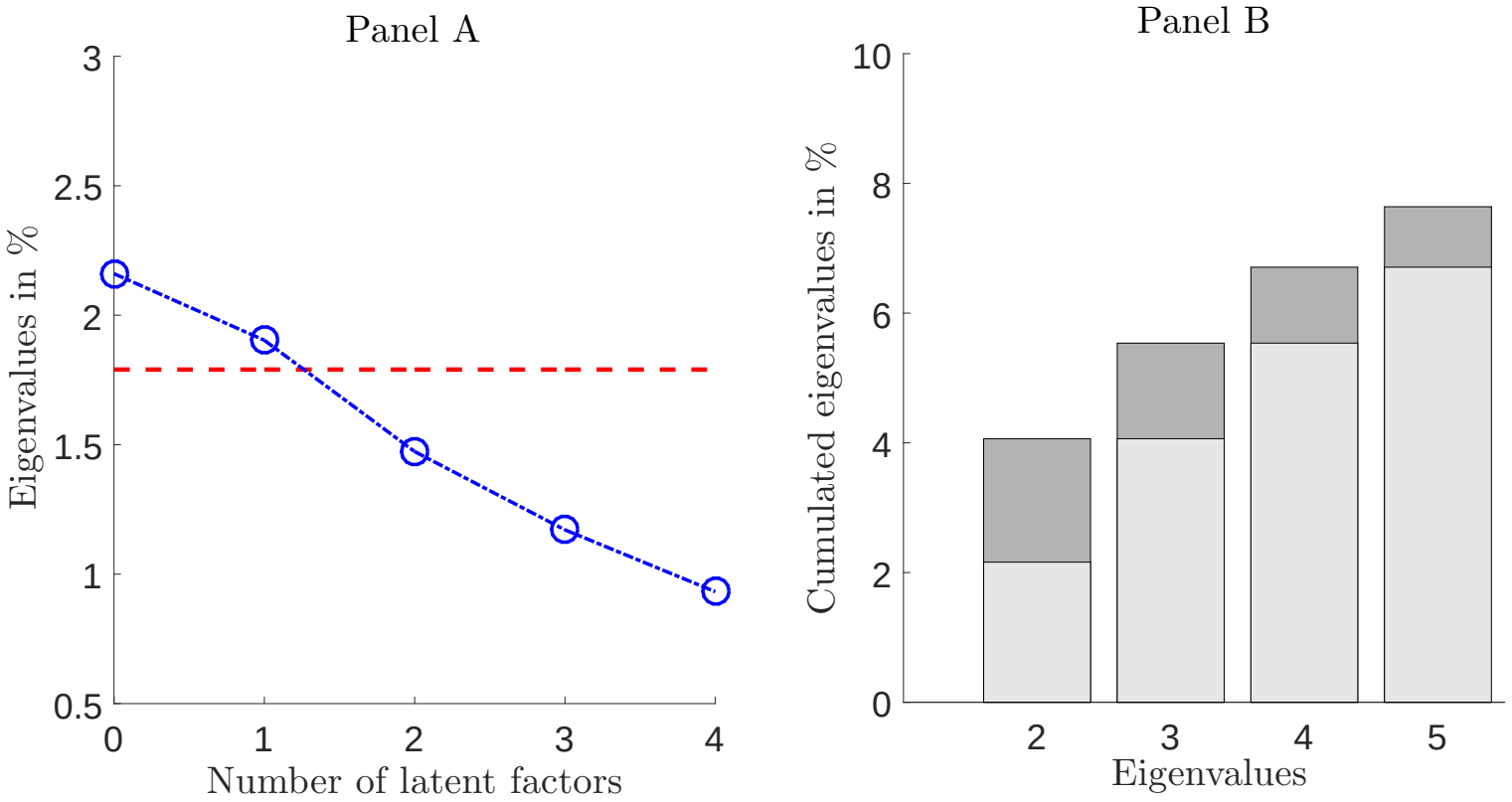
where $n^{\chi} = \sum_i \mathbf{1}_i^{\chi}$ and the **eigenvalues** are interpreted as **percentage of the variance** of the normalised residuals.

We also investigate the ratio

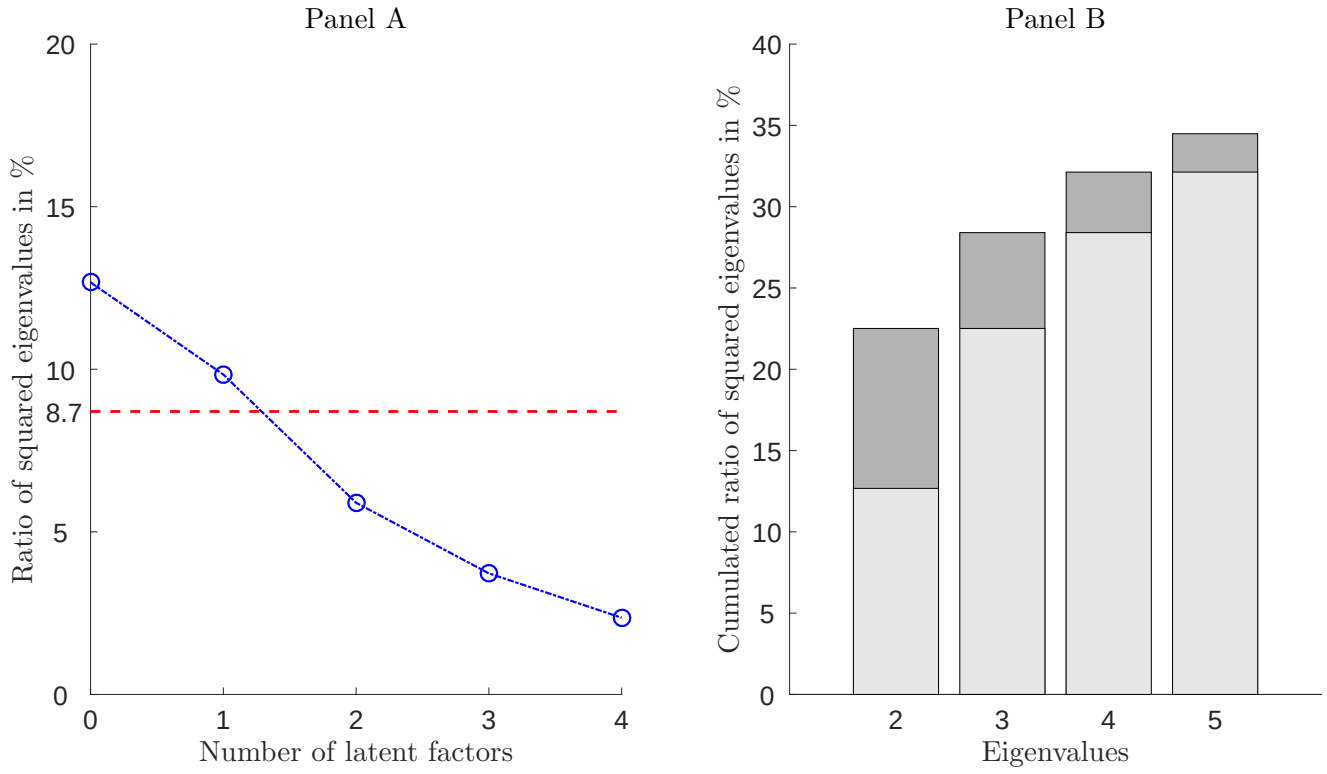
$$\mu_j^2 \left(\frac{1}{nT} \sum_i \mathbf{1}_i^{\chi} \bar{\bar{\varepsilon}}_i \bar{\bar{\varepsilon}}_i' \right) / \sum_{l=1}^{T} \mu_l^2 \left(\frac{1}{nT} \sum_i \mathbf{1}_i^{\chi} \bar{\bar{\varepsilon}}_i \bar{\bar{\varepsilon}}_i' \right).$$

- This quantity is a measure the contributions of the omitted factors in terms of the **off-diagonal terms** (correlation part) in addition to the diagonal terms (residual variance).

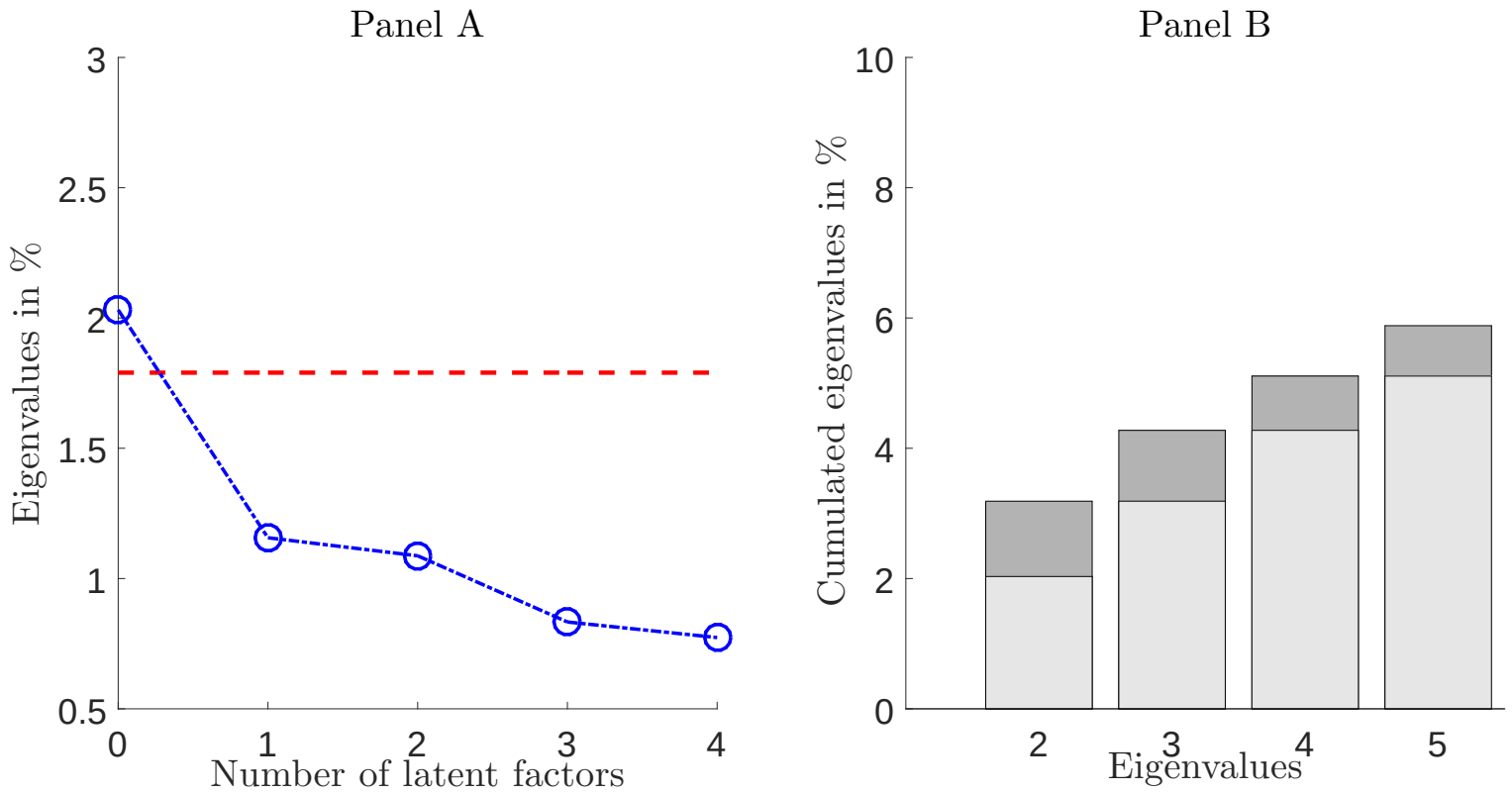
Time-invariant CAPM



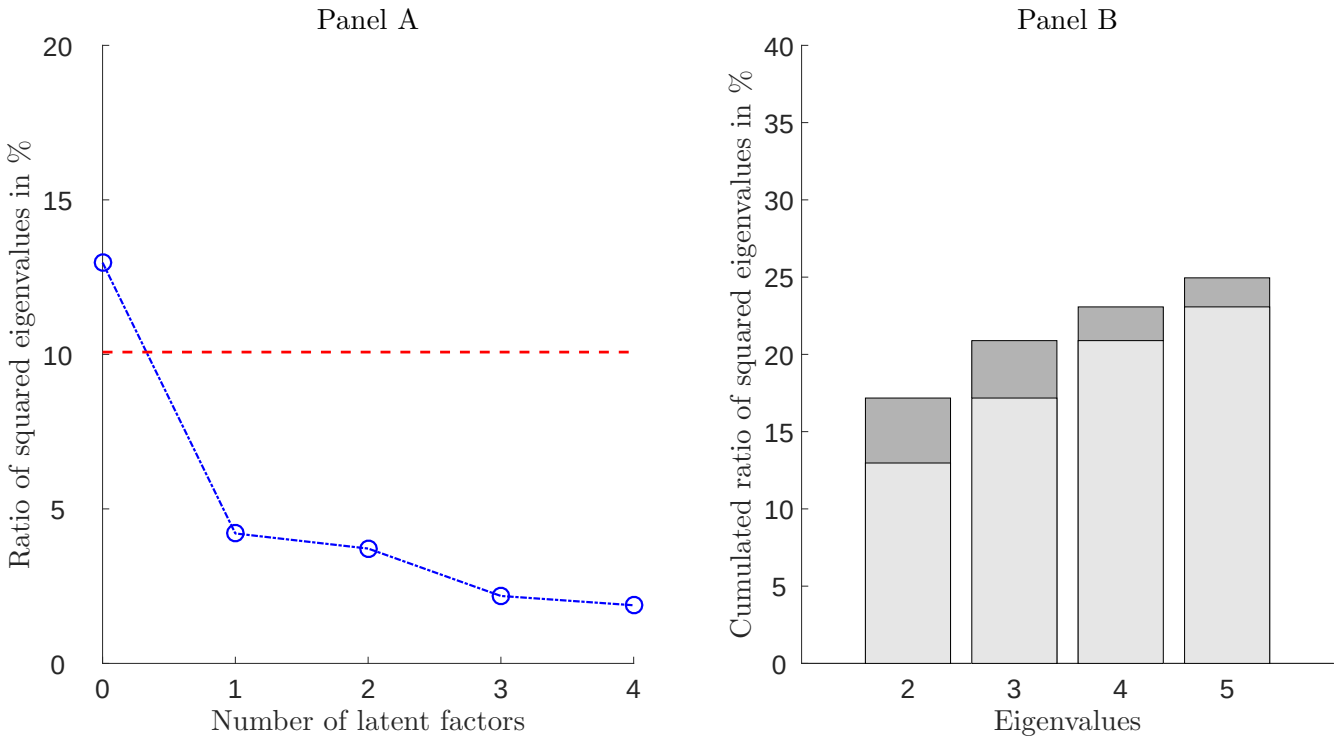
Time-invariant CAPM



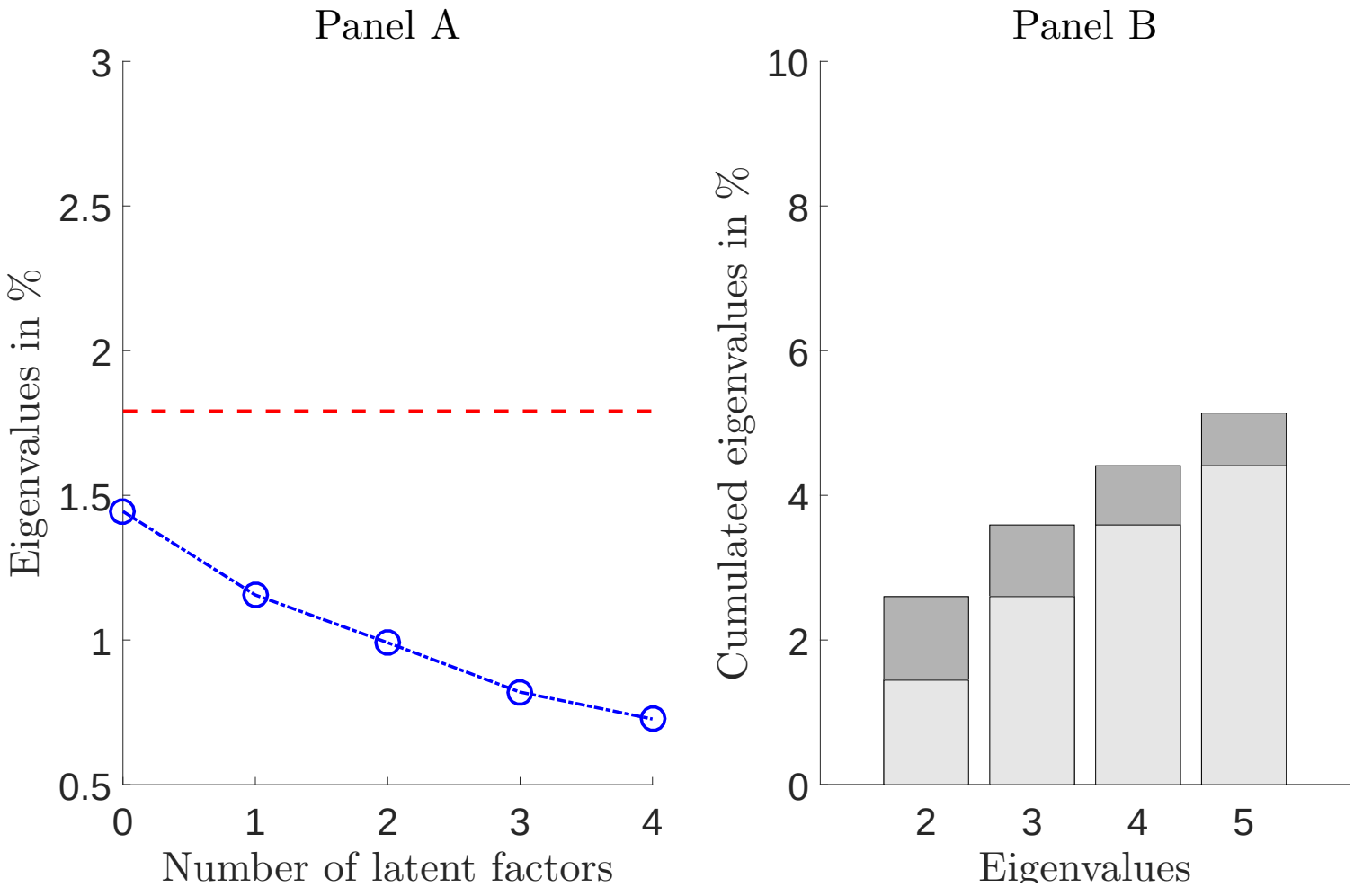
Time-invariant three-factor Fama-French model



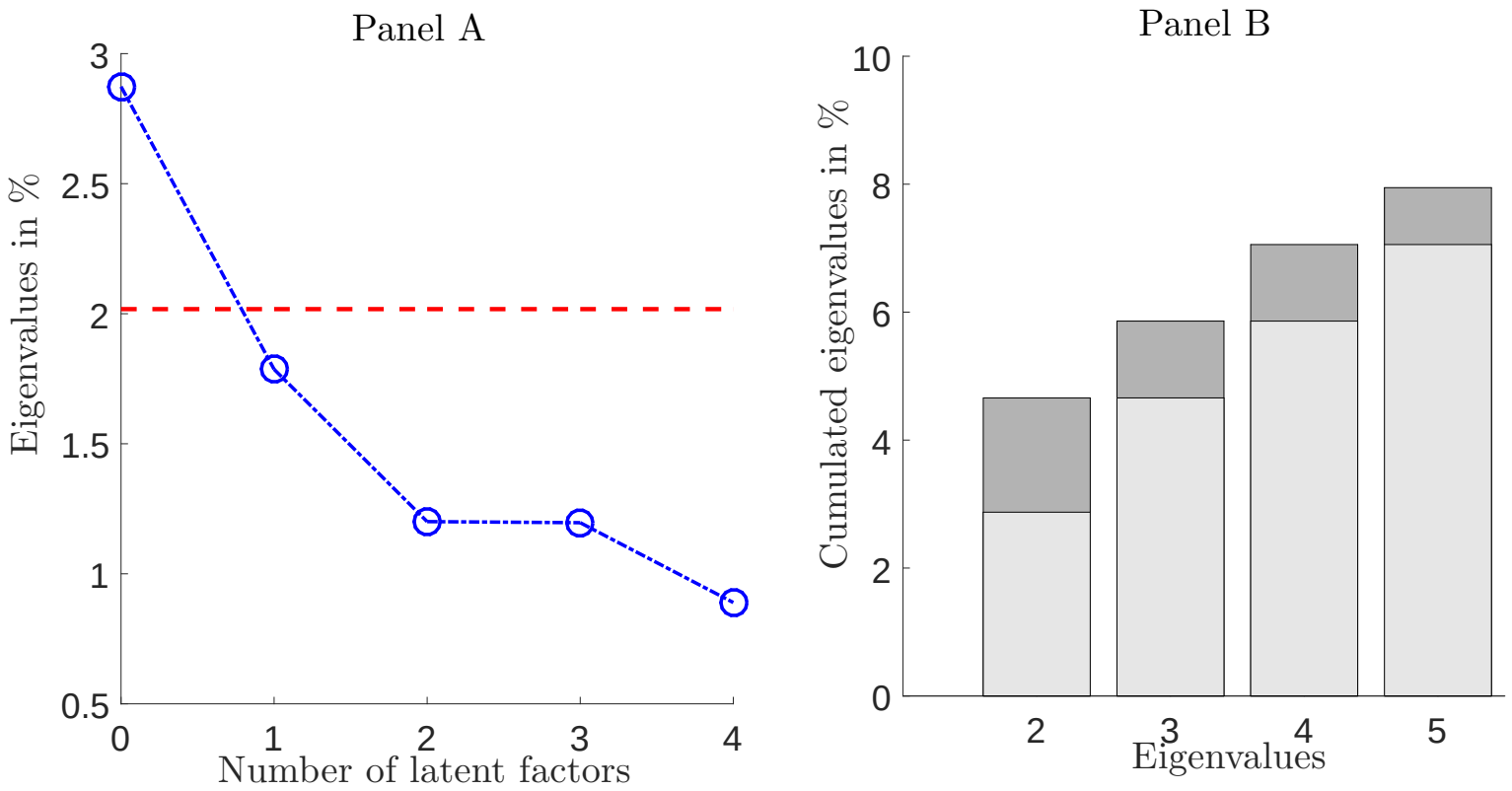
Time-invariant three-factor Fama-French model



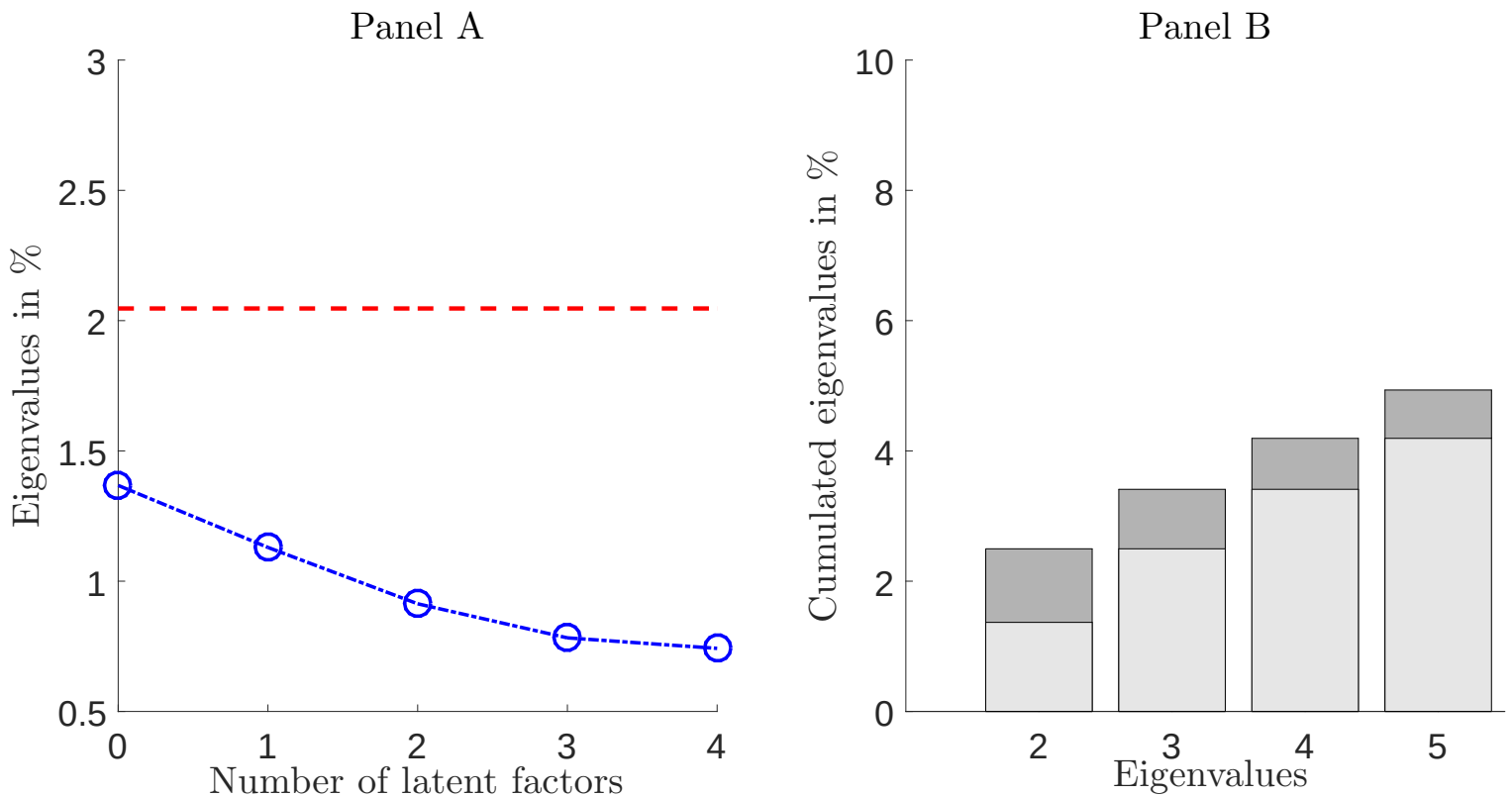
Time-invariant five-factor Fama-French model



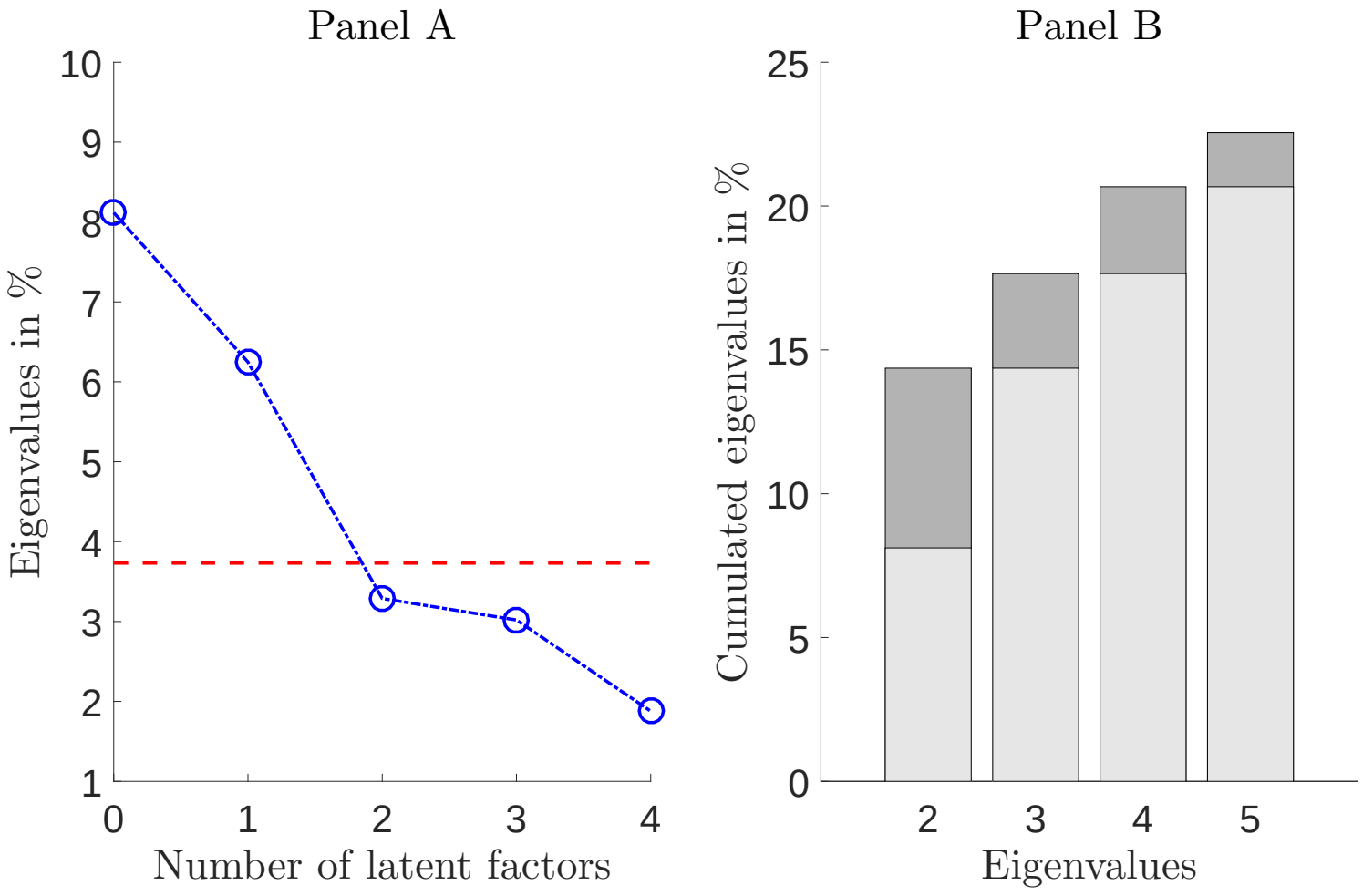
Time-varying CAPM with $Z_{t-1} = (1, divY_{t-1})'$



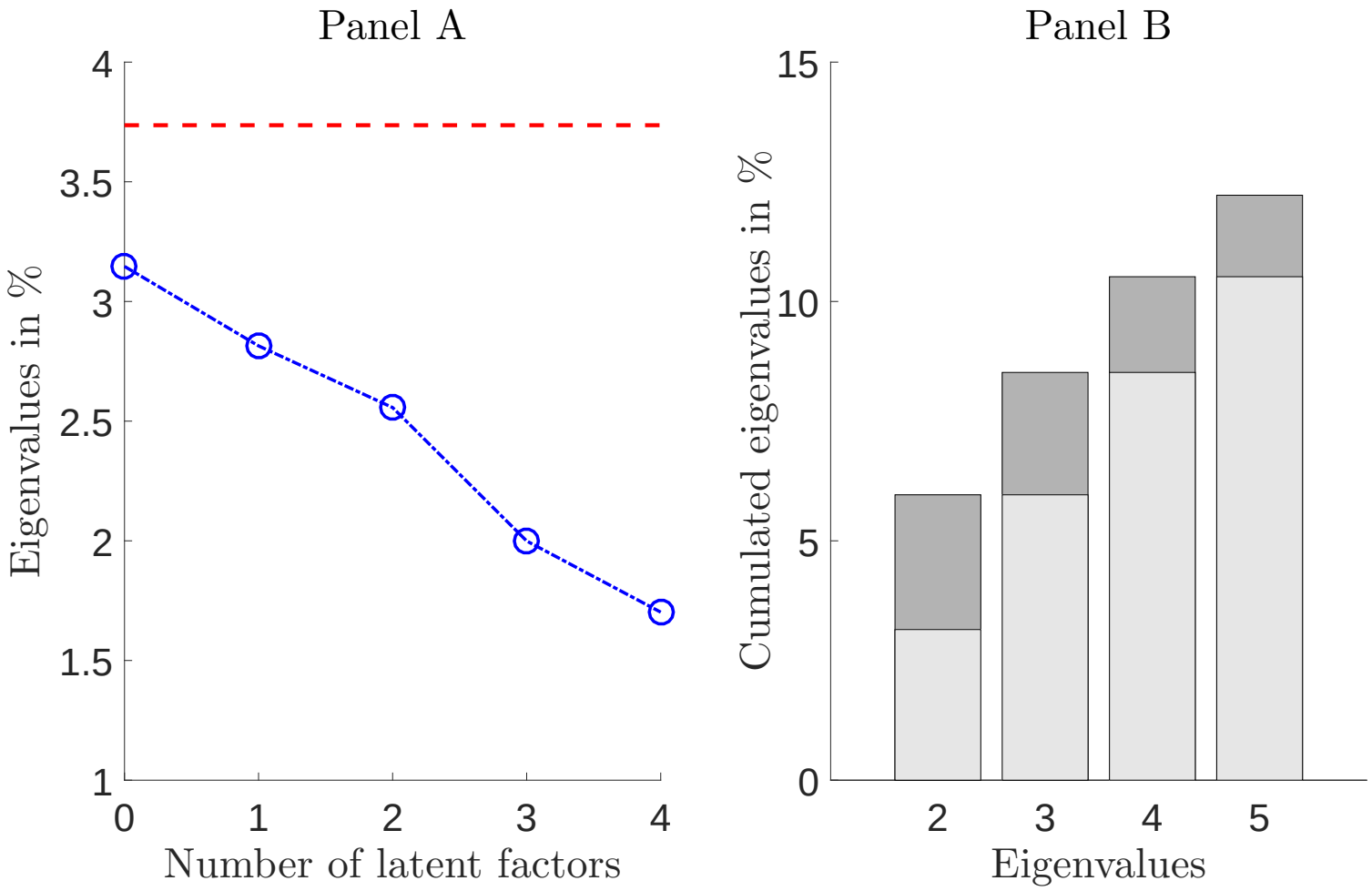
Time-varying three-factor Fama-French model with $Z_{t-1} = (1, divY_{t-1})'$



Time-invariant CCAPM



Time-invariant Epstein-Zin model



Conclusions

- A new diagnostic criterion for approximate factor structure in large cross-sectional datasets
- The simple criterion is based on three steps:
  (i) compute the largest eigenvalue of a variance-covariance matrix,
  (ii) substract a penalty,
  (iii) conclude on the validity of the approximate factor structure if criterion value is negative.
- The theoretical results are obtained on residuals, instead of true errors, allowing for unbalanced panel and considering an asymptotics with $n >> T$.

Empirical results:

- interpretation of the diagnostic criterion as the percentage of the residuals' variance explained by omitted latent factors;
- interpretation of the squared eigenvalues as the percentage of the correlation part explained by omitted latent factors (see Fiorentini-Sentana, 2015);
- we can choose either among time-invariant specifications with at least four financial factors, or a scaled Fama-French model;
- the latent factors are more representative of the correlation part than the variance part of the residuals;
- we cannot select macroeconomic models without the market factor.