MS&E 349: Homework 2

Markus Pelger

Due 5pm May 26th 2017, Submit by email

Please submit one homework assignment per group. The homework solution should be submitted online by email to mpelger@stanford.edu. Please indicate clearly the names of all group members. The solutions should be typed in Latex and include the commented code in an appendix section. Please also submit the executable and commented code.

Part 1: Factor Modeling

Question 2 and 3 require essentially the same code but apply it to different data sets.

Question 1 Theory

Consider a panel

$$y_{i,t} = \beta^\top x_{i,t} + \epsilon_{i,t} \qquad i = 1, ..., N \;\; t = 1, ..., T$$

Assume that $x_{i,t}$ does not vary over time and that $\epsilon_{i,t}$ is independent over time. Show that the pure cross-sectional regression

$$E_T[y_t] = E_T[x]\beta + E_T[\epsilon_t]$$

is identical to the Fama-MacBeth regression.

. . . . . .

Question 2 Simulation

We simulate excess returns of N assets over T time periods following an approximate factor model with 3 factors:

$$X_{t,i} = F_t \Lambda_i^\top + e_{t,i} \qquad i = 1, ..., N \;\; t = 1, ..., T$$

In matrix notation this reads as

$$\underbrace{X}_{T\times N} = \underbrace{F}_{T\times K}\,\underbrace{\Lambda^\top}_{K\times N} + \underbrace{e}_{T\times N}$$

with $K = 3$. Our goal is to estimate the unknown latent factors F and the loadings Λ.

The parameters are chosen to create a realistic three factor model where the first factor is the market and the other two factors are weaker. One of the weaker factors has a high Sharpe-ratio representing for example a value factor. One of the weak factors has a small Sharpe ratio approximating an industry factor.

Factors:

- 1. Factor represent the market with i.i.d. returns modeled as $N(1.2, 9)$ and hence a Sharpe-ratio of 0.4
- 2. Factor represents an industry factors following i.i.d. $N(0.1, 1)$ and has a Sharpe-ratio of 0.1.

- 3. Factor has a small variance but high Sharpe-ratio. Its excess-returns follow an i.i.d. $N(0.4, 0.16)$ and thus have a Sharpe-ratio of 1.

The loadings are normalized such that $\frac{1}{N}\Lambda^\top\Lambda = I_4$. Intuitively all three factors affect all N assets equally strong The first loading vector is $\mathbb{1}$, i.e. it represents a market factor that affects all assets equally. The other two loading vectors are i.i.d. draws from a normal distribution $N(0, 1)$. It assumes that the portfolio weights of the second and third factor are random with a variance of 1. As the loadings are normalized the strength of the factors is completely determined by their variance and mean. We could just as well normalize the variances of the factors and adjust the loadings and portfolio weights. This means that a factor with a small variance can be interpreted as a factor that only affects a smaller number of assets.

We allow for cross-sectional and time-series correlation and heteroskedasticity in the residuals:

$$e = \sigma_e D_T A_T \epsilon A_N D_N$$

Errors:

- ϵ is a $T \times N$ matrix and follows a multivariate standard normal distribution
- Time-series correlation in errors: A_T creates an AR(1) model with parameter ρ.
- Cross-sectional correlation in errors: A_N is a Toeplitz-matrix with $(\beta, \beta, \beta, \beta^2)$ on the right four off-diagonals with parameter β.
- Cross-sectional heteroskedasticity: D_N is a diagonal matrix with independent elements following $N(1, 0.2)$
- Time-series heteroskedasticity: D_T is a diagonal matrix with independent elements following $N(1, 0.2)$
- Signal-to-noise ratio: σ_e^2

We set the number of observations to $N = 100$ and $T = 150$.
Please simulate 3 different data sets:

1. Toy model with i.i.d. residuals: $\rho = 0$, $\beta = 0$, $D_N = I_N$, $D_T = I_T$, $\sigma_e^2 = 1$
2. Toy model with i.i.d. residuals and large signal-to-noise ratio: $\rho = 0$, $\beta = 0$, $D_N = I_N$, $D_T = I_T$, $\sigma_e^2 = 25$.
3. Realistic model: $\rho = 0.1$, $\beta = 0.7$, D_N and D_T with diagonal entries following i.i.d. $N(1, 0.2)$, $\sigma_e^2 = 10$

For each of the 3 data sets answer the following questions:

1. Estimate the number of latent factors based on the sample covariance matrix with the Bai and Ng (2002) information criteria. Choose an appropriate penalty function. Plot the criteria function against the number of factors (for $k = 1, ..., 15$). Report your estimate.

2. Estimate the number of latent factors based on the sample covariance matrix with the Ahn and Horenstein (2013) eigenvalue ratio test. Plot the eigenvalue ratios against the number of factors (for $k = 1, ..., 15$). Report your estimate.

3. Estimate the number of latent factors based on the sample covariance matrix with the Onatski (2010) eigenvalue difference test. Choose an appropriate cutoff value. Plot the eigenvalue differences against the number of factors (for $k = 1, ..., 15$). Report your estimate.

4. Estimate 3 factors and corresponding loadings with PCA of the sample covariance matrix. Denote this estimator by $\hat{F}^{\text{PCA}}$.

5. Estimate 3 factors and corresponding loadings with a new method called Risk-Premium PCA. Denote this estimator by $\hat{F}^{\text{RP-PCA}}$.

   Conventional PCA tries to explain as much variation as possible. Conventional statistical factor analysis applies PCA to the sample covariance matrix $\frac{1}{T}X^\top X - \bar{X}\bar{X}^\top$ where $\bar{X}$ denotes the sample mean of excess returns. The eigenvectors of the largest eigenvalues are proportional to the loadings $\hat{\Lambda}^{\text{PCA}}$. Factors are obtained from a regression on the estimated loadings. It can be shown that conventional PCA factor estimates are based on the variation objective function: [1]

$$\min_{\Lambda,F} \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ti} - F_t\Lambda_i^\top)^2$$

   The new approach is Risk-Premium-PCA (RP-PCA). It applies PCA to a covariance matrix with overweighted mean

$$\frac{1}{T}X^\top X + \gamma\bar{X}\bar{X}^\top$$

   with the risk-premium weight γ. The eigenvectors of the largest eigenvalues are proportional to the loadings $\hat{\Lambda}^{\text{RP-PCA}}$. RP-PCA minimizes jointly the unexplained variation and pricing error:

$$\min_{\Lambda,F} \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ti} - F_t\Lambda_i^\top)^2}_{\text{unexplained variation}} + \gamma \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_i - \bar{F}\Lambda_i^\top\right)^2}_{\text{pricing error}}$$

   where $\bar{F}$ denotes the sample mean of the factors. The factors $\hat{F}^{\text{RP-PCA}}$ are estimated by a

[1] The variation objective function assumes that the data has been demeaned.

regression of the asset returns on the loadings $\hat{\Lambda}^{\text{RP-PCA}}$.
Choose a risk-premium value of $\gamma = 50$ for your estimation.

6. Calculate the maximum Sharpe-ratio than can be obtained by the factors. Do this for the true factors, the PCA and the RP-PCA factors. Hint: The maximum Sharpe-ratio that we can obtain as a linear combination of the factors equals $\sqrt{\mu_F^\top \Sigma_F^{-1} \mu_F}$ and measures how well the factors can approximate the stochastic discount factor. (μ_F is the factor mean and Σ_F the factor covariance matrix.)

7. Calculate the estimation error between the true and estimated factors for PCA and RP-PCA. For each true factor run a regression on the three estimated factors. This fit gives you the right rotation of the estimated factors. Then use the fitted estimated factors to calculated the root-mean-square error for each factor: $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(F_{t,k} - \hat{F}_{t,k})^2}$

8. For each factor plot the true and the two estimated sample paths. Plot the cumulative return series instead of the return series.

9. Run a time-series asset pricing test for the first 20 assets for the PCA and RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

10. Run a time-series asset pricing test for all the assets for the PCA and RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

11. Run a cross-sectional asset pricing test for the first 20 assets for the PCA and RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

12. Run a cross-sectional asset pricing test for all the assets for the PCA and RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

13. Plot expected and predicted excess returns for each portfolio as a scatterplot and fit a straight line through the points.

14. Comment on the major findings from the previous subquestions.

. . . . . .

Question 3 Empirical

Download three different sorted portfolio data sets from Kenneth French's website:
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

1. 25 value-weighted size and book-to-market sorted portfolios, monthly returns from 1963/07-2017/01

2. 25 value-weighted size and accrual sorted portfolios, monthly returns from 1963/07-2017/01

3. 49 industry portfolios, monthly returns from 1963/07-2017/01

Download also the monthly 3 Fama-French factors for the same time period.
For each of the 3 data sets answer the following questions:

1. Estimate the number of latent factors based on the sample covariance matrix with the Bai and Ng (2002) information criteria. Choose an appropriate penalty function. Plot the criteria function against the number of factors (for $k = 1, ..., 15$). Report your estimate.

2. Estimate the number of latent factors based on the sample covariance matrix with the Ahn and Horenstein (2013) eigenvalue ratio test. Plot the eigenvalue ratios against the number of factors (for $k = 1, ..., 15$). Report your estimate.

3. Estimate the number of latent factors based on the sample covariance matrix with the Onatski (2010) eigenvalue difference test. Choose an appropriate cutoff value. Plot the eigenvalue differences against the number of factors (for $k = 1, ..., 15$). Report your estimate.

4. Estimate 3 factors and corresponding loadings with PCA of the sample covariance matrix. You will get 3 statistical factors for each of the three sets of portfolios. Only use the statistical factors estimated for the specific data set for the next subquestions.

5. Estimate 3 factors and corresponding loadings with Risk-Premium PCA. Choose a risk-premium value of $\gamma = 100$ for your estimation. You will get 3 statistical factors for each of the three sets of portfolios. Only use the statistical factors estimated for the specific data set for the next subquestions.

6. Plot for each PCA and each RP-PCA factor the loadings for all stocks (in 6 separate plots). Interpret your results.

7. Calculate the maximum Sharpe-ratio for the market factor, the 3 Fama-French factors, the 3 PCA and 3 RP-PCA factors.

8. Run a time-series asset pricing test for the market factor, the 3 Fama-French factors, the 3 PCA and 3 RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

9. Run a cross-sectional asset pricing test for the market factor, the 3 Fama-French factors, the 3 PCA and 3 RP-PCA factors. You can use the test-statistic that assumes no time-series correlation.

10. Plot expected and predicted excess returns for each portfolio as a scatterplot and fit a straight line through the points. Do this for each set of factors.

11. Comment on the major findings from the previous subquestions.

. . . . . .

Part 2: High-Frequency

Please load the data set "HF_Data.csv". It contains high-frequency data for the SPY ETF tracking the S&P500 index (called the market return from here forward) and for Apple for 2012. The data are log price increments based on 5 minutes data with the first observation at 9:35am and the last at 4pm. Each day includes 77 log-price increments. The data is already cleaned. You could download it from the TAQ data base on WRDS. The file also contains daily returns for the SPY index and Apple.

Question 4 Empirical

1. Calculate weekly volatilities (i.e based on 385 observations) and plot them for the SPY index and Apple. The weekly "spot volatility" is the quadratic covariation for one week.
2. Estimate the jumps for the SPY index and Apple. Justify your approach.
3. Calculate the weekly volatilities based on the continuous data, i.e. after removing the jumps. Plot them. Are your estimates different from the previous approach?
4. Calculate the weekly jump variation process for the SPY and Apple. Plot them.
5. Calculate the weekly market beta for Apple using all price observations (i.e. including the jumps). Plot them.
6. Calculate the weekly market beta for Apple using only continuous price movements. Plot them. Are they different than those including the jumps?
7. Calculate yearly market betas for Apple using all high-frequency data, only the continuous movements, only the jump movements or only the daily returns. Comment on your results.
8. Do you think that market betas are time-varying? Justify your answer.

. . . . . .