Factor Models 1: The Static Factor Model and Principal Components

Advanced Time Series Econometrics

Scottish Graduate Programme in Economics

Introduction

- Big Data can be Fat or Tall
- Tall Data = huge number of observations (e.g. billions of internet searches or supermarket purchases)
- Fat Data = huge number of variables
- Factor models are popular method for working with Fat Data
- E.g. many assets which can make up financial portfolio
- E.g. many macroeconomic variables to help forecast GDP growth
- This lecture will cover both static and dynamic factor models
- Static factor models popular in finance
- Dynamic factor models (DFMs) popular in macroeconomics

Readings

- For static factor models, course textbook offers imperfect coverage
- Ghysels and Marcellino (G+M) textbook has good coverage of dynamic factor models (next week's lecture)
- G+M do not cover static factor models
- For static factor modes Tsay, Chapter 9 is good
- To replicate some of the exercises, see the **R** code available on Tsay's website: faculty.chicagobooth.edu/ruey-s-tsay/research/multivariate-time-series-analysis-with-r-and-financial-applications
- Lecture slides also can count as "reading" a reading for these topics

Why Does Fat Data Arise in Macroeconomics and Finance?

- Financial data: hundreds or thousands of assets can go into financial portfolio
- Macroeconomic data sets often involve hundreds of variables
- US: FRED data bases (monthly and quarterly) for hundreds of variables
- https://research.stlouisfed.org/econ/mccracken/fred-databases/
- More information is better, therefore why not use them all?
- Avoid risk of mis-specification (omitted variables bias)
- Many other countries have similar data sets

Other Place Fat Data Arises

- Data for many countries or regions within a country
- E.g. even with 10 variables for 20 countries want to work with model for 200 variables
- Macro models which combine macro and financial variables and different frequency (covered in later lecture)
- Disaggregation (e.g. components of inflation: food, durable goods, etc. etc. or industrial sectors)
- Term structure (e.g. bonds of different maturity: overnight, 3 month, 1 year, 5 year, 10 year)
- Several assets, each with different maturity can easily end up with lots of variables
- Internet search data, supermarket scanner data, etc. etc.
- Many cases where we have lots of variables, how best to extract the information out of them in an econometric model?
- Factor models are a popular choice

The Static Factor Model

- Commonly used in finance
- Returns of financial assets tend to have little persistence
- E.g. in AR(1) model tend to find ρ near zero
- Returns tend to be unpredictable (or nearly so), if not it would be easy to make lots of money
- In contrast macroeconomic variables have more persistence
- E.g. GDP growth data from last lecture had $\rho = 0.37$
- For macro data Dynamic Factor Model more appropriate

The Static Factor Model

- r_{it} is the return on asset i in period t
- $i = 1, .., N$
- $t = 1, .., T$
- N can be large
- Static factor model is:

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + ... + \beta_{im} f_{mt} + \varepsilon_{it}$$

- Return of each asset depends on m factors: $f_{1t} ... f_{mt}$
- m is typically small
- Fat Data problem turns into N individual small regressions
- Notation: $\beta_{i1}, .., \beta_{im}$ are called *factor loadings*
- ε_{it} called idiosyncratic component or asset specific factor

The Static Factor Model

- ε_{it} is error specific to asset i
- $cov\left(\varepsilon_{it}, \varepsilon_{jt}\right) = 0$ for $i \neq j$
- $cov\left(\varepsilon_{it}, f_{jt}\right) = 0$ for all $i \neq j$
- $var\left(\varepsilon_{it}\right) = \sigma_i^2$
- Errors uncorrelated over time (white noise)
- So far Static factor model is just a set of regressions (one for each asset)

What are the Factors?

- Different models have different treatment of factors
- We will go through several models
- Idea: factors contain information that is common to all assets
- $f_t = \begin{pmatrix} f_{1t} \\ . \\ . \\ f_{mt} \end{pmatrix}$ sometimes called *common factors*
- Idea: asset returns are driven by these common factors
- These common factors influence each asset differently (factor loadings different for each asset)
- Then add asset specific factor (ε_{it})

Assumptions About the Factors

- In general, factors are unobserved random variables (similar to states in state space model)
- In some cases, factors are observed variables
- If random variables assume

$$E\left(f_t\right) = \mu_f$$

$$cov\left(f_t\right) = \Sigma_f$$

Summary of Static Factor Model in Matrix Notation

- Model can be written as:

$$r_t = \alpha + \beta f_t + \varepsilon_t$$

- r_t is $N \times 1$ vector of asset returns
- f_t is $m \times 1$ vector of factors
- β is $N \times m$ matrix of factor loadings
- α is $N \times 1$ vector of intercepts
- ε_t is $N \times 1$ vector of mean zero asset specific factors

Summary of Static Factor Model in Matrix Notation

- $$cov\left(\varepsilon_t\right) = D$$

  where D is diagonal matrix with σ_i^2 for $i = 1, .., N$ on diagonal
- As for the asset returns themselves can easily show

$$cov\left(r_t\right) = \beta \Sigma_f \beta' + D$$

- Econometric estimation: if factors are observed least squares methods can be used
- Since D is diagonal OLS equation by equation can also be used
- Each equation is a multivariate regression and the `lm()` command in **R** can be used

The Market Model

- Classic model (Sharpe, 1970) which assumes one known factor
- Factor is the "market return" = measure of returns on stock market as a whole
- Model is

$$r_{it} = \alpha_i + \beta_i r_t^m + \varepsilon_{it}$$

- r_t^m is excess returns on the stock market (e.g. FTSE or SP500)

What is Market Model used For?

- Financial economists use terminology arising from capital asset pricing model (CAPM):
- "the alpha and beta of a particular stock"
- The CAPM alpha and beta are the intercept and factor loading in the market model
- These are estimated using the market model and useful for financial practitioner
- alpha measure of whether a stock is out-performing market (on risk adjusted basis)
- alpha is expected to be zero
- beta is measure of whether asset more/less volatile/risky than market as a whole
- Market portfolio (e.g. S&P500) has beta of 1
- $\beta < 1$ means less volatile, $\beta > 1$ more volatile

What is Market Model used For?

- Portfolio choice involves:
- Choose weights (shares), $s = (s_1, .., s_N)'$, attached to each of N assets
- Weights sum to one, but some may be negative (short selling)
- Global minimum variance portfolio (GMVP) chooses weights to minimize risk
- For market model, can work out GMVP weights to be

$$s = \frac{(\beta \Sigma_f \beta' + D)^{-1} \iota}{\iota' (\beta \Sigma_f \beta' + D)^{-1} \iota}$$

- ι is vector of ones
- Market model gives us estimates of everything in formula for s and, thus, to calculate GMVP

Example: The Market Model

- Monthly data, Jan 1990 to Dec 2003 on excess return on stocks in 13 companies
- Excess return = return over risk free rate (3 month T-bill rate)
- Abbreviations for companies are: AA, AGE, CAT, F, FDX, GM, HP, KMB, MEL, NYT, PG, TRB, TXN
- These are dependent variables
- For excess market return r_t^m use return on S&P500
- This is market factor (explanatory variable)
- Table on next slide contains results using OLS
- P_α is p-value for testing $H_0 : \alpha = 0$
- P_β is p-value for testing $H_0 : \beta = 0$

Estimates from the Market Model						
	$\widehat{\alpha}_i$	P_α	$\widehat{\beta}_i$	P_β	R_i^2	s_i
AA	0.54	0.36	1.29	0.00	0.35	0.12
AGE	0.72	0.24	1.51	0.00	0.42	-0.03
CAT	0.84	0.16	0.94	0.00	0.22	0.08
F	0.45	0.48	1.22	0.00	0.29	0.02
FDX	0.80	0.25	0.81	0.00	0.14	0.08
GM	0.20	0.75	1.05	0.00	0.24	0.05
HPQ	0.68	0.35	1.63	0.00	0.36	-0.04
KMB	0.55	0.25	0.55	0.00	0.13	0.25
MEL	0.88	0.06	1.12	0.00	0.39	0.07
NYT	0.49	0.34	0.77	0.00	0.21	0.15
PG	0.89	0.08	0.47	0.00	0.09	0.24
TRB	0.65	0.25	0.72	0.00	0.16	0.14
TXN	1.44	0.11	1.80	0.00	0.32	-0.04

Example: The Market Model

- Table presents alphas and betas for each company
- In table no α_i is significantly different from zero at 5% level
- β_i for some companies below one, for others above
- GMVP weights spread over stocks, but roughly half weight in two companies: KMB and PG
- Some stocks have negative weights (short-selling)
- Useful information for financial practitioners
- R_i^2 tend to be fairly low (market return only explains a small part of variation in each stock's price)
- A great deal of heterogeneity across companies

Static Factor Model: Other Models with Known Factors

- Several other models which use observed data as factors
- Market model use one variable (market return)
- Multifactor models use more variables
- E.g. Chen, Roll and Ross (1986) use unexpected changes (surprises) in inflation and unemployment as two factors
- Unexpected change = first estimate a model (e.g. a VAR), fitted values tell you what is expected
- Difference between actual data and fitted values are unexpected
- Fundamental factor models: asset specific variables (e.g. market capitalization, book value, industrial sector as factors)

Static Factor Model: Unobserved Factors

- But what if you do not know what the factors are?
- E.g. you do not know that the one thing common to assets for all companies is the market rate?
- You just want to assume there is something common to all assets without specifying exactly what it is
- Most popular way of estimating factors is principal components analysis (PCA)
- Other methods of treating factors as unobserved (discussed later with dynamic factor models)

Principal Components

- Forget about the static factor model for now
- Return to it shortly
- Preview of later result: Principal components (PC) can be used as estimates of factors
- But first explain ideas of principal components analysis (PCA)
- Remember r_t is $N \times 1$ vector of asset returns
- Let Σ_r be covariance matrix (large $N \times N$ matrix)
- Perhaps a small number of linear combinations of r contain most of the variation in r?
- This is idea of PCA: squeeze all the variability in N assets into a few components (factors)

Principal Components: Theory

- $w_i = N$ vector of weights
- w_{i1} = weight attached to asset 1, etc.
- Normalize: $w_i' w_i = 1$
- Define:

$$\begin{aligned} f_1 &= w_1' r \\ f_2 &= w_2' r \\ & \quad etc. \end{aligned}$$

- f_1 is $T \times 1$ vector, weighted average of all the asset returns using weights w_1
- f_2 is $T \times 1$ vector, weighted average of all the asset returns using weights w_2
- etc.

Principal Components: Theory

- PCA wants to find weights such that:
- $f_1, f_2, ...$ are uncorrelated with one another
- their variances are as large as possible
- Intuition: variance = variability = information useful for explaining a dependent variable
- E.g. remember in regression want to have explanatory variables with as high a variance as possible
- Solution to PCA problem: maximize something subject to some constraint
- Calculus problem, I will not provide proof (see Tsay page 484)
- Involves eigenvalues and eigenvectors of Σ_r

Example: PCA

- Monthly data (Jan 1990 through December 2008) for returns on 5 stocks:
- IBM, HPQ, INTC, JPM and BAC
- In **R**: Run the PCA command `prcomp()`
- Table on next slide contains some of the output produced

PCA Output Relating to Variation Explained by Factors		
Component/Factor	Eigenvalue	Proportion
1	2.61	0.52
2	1.07	0.21
3	0.57	0.11
4	0.45	0.09
5	0.30	0.06

Interpretation of PCA Results

- How can previous table be interpreted?
- Σ_r is 5×5 matrix so has 5 eigenvalues
- Do not worry if you do not know what an eigenvalue is, key column is "Proportion"
- Note column labelled Proportion is proportional to Eigenvalue
- Proportion = proportion of total variability in data explained by factors
- f_1 contains 52% of the total variability in r_t
- f_2 contains 21% of the total variability in r_t
- etc.

Interpretation of PCA Results

- But what are $f_1, f_2, \ldots$?
- Weighted average of all returns for all the assets
- But what are the weights?
- These are the eigenvectors produced by the `prcomp()` command
- See table on next slide

Eigenvectors (Weights Used in the Factors)					
Variable	f_1	f_2	f_3	f_4	f_5
IBM	0.43	0.34	0.84	0.00	0.01
HPQ	0.46	0.36	-0.38	0.70	0.15
INTC	0.45	0.39	-0.39	-0.70	0.02
JPM	0.48	-0.47	-0.05	0.05	-0.74
BAC	0.42	-0.62	0.04	-0.07	0.66

Interpretation of PCA Results

- We now know that:

$$f_{1t} = 0.43r_t^{IBM} + 0.46r_t^{HPQ} + 0.45r_t^{INTC} + 0.48r_t^{JPM} + 0.42r_t^{BAC}$$

- First factor (which explains over 50% of total variability) is approx. equally weighted average of returns of all assets
- Tsay calls this "market component" (average over whole market)
- Second factor (explains 21%):

$$f_{2t} = 0.34r_t^{IBM} + 0.36r_t^{HPQ} + 0.39r_t^{INTC} - 0.47r_t^{JPM} - 0.62r_t^{BAC}$$

- IBM, HPQ and INTC are computer companies, JPM and BAC are banks
- Second factor (roughly) takes difference between average returns to computer companies and banks
- Tsay calls this "industrial component"

Using PCA in the Static Factor Model

- Key point for empirical practice:
- Statistical packages in **R** do PCA for you
- Provide estimates of $f_1, f_2 \ldots$
- These can be used as factors in the static factor model
- Treat like explanatory variables and use OLS
- Instead of having N explanatory variables, reduced to m factors
- A few factors can have most of the variability in entire set of N variables
- Useful in many Big Data contexts

Selecting the Number of Factors

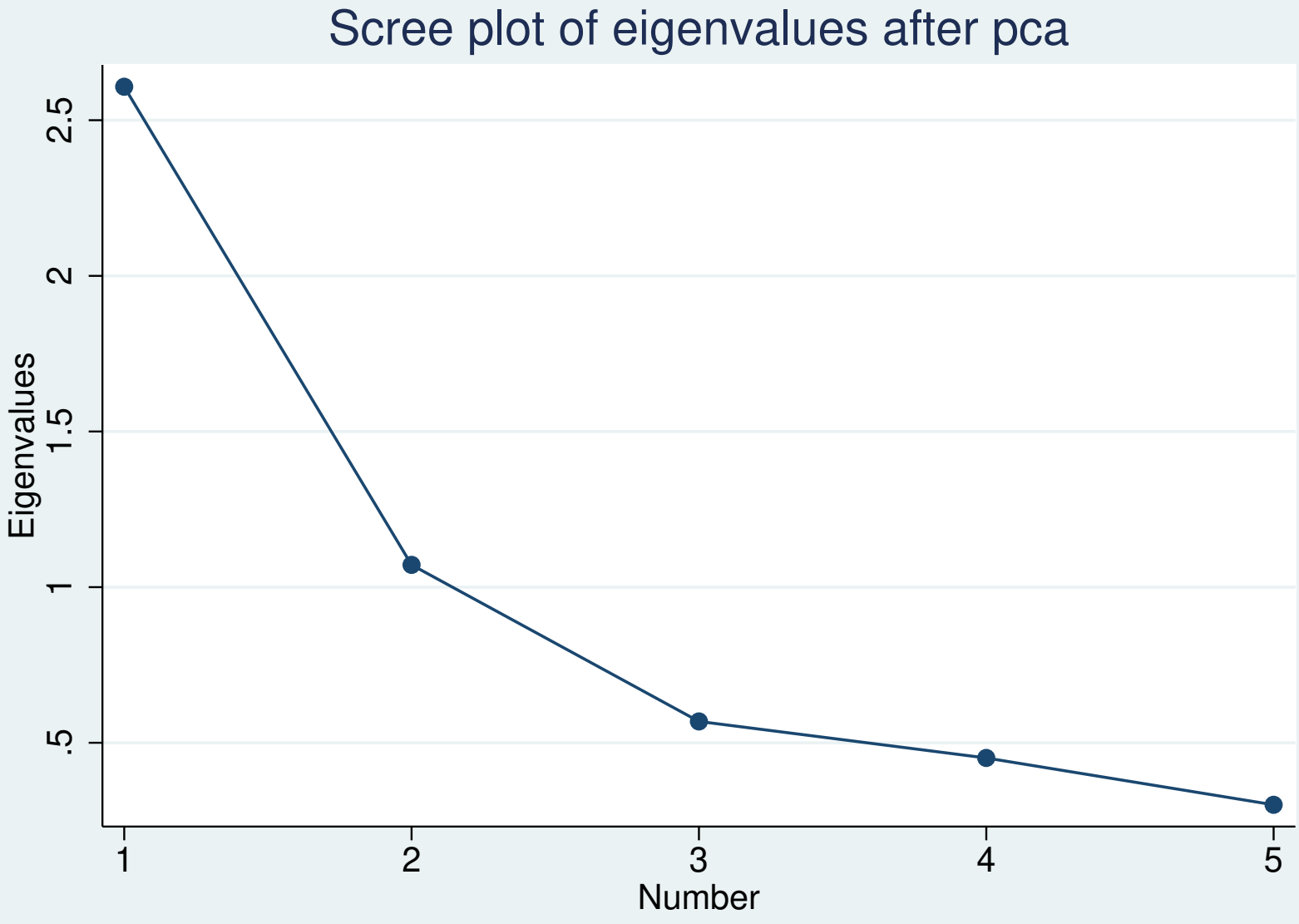
- Several ways for selecting m
- Informal: choose factors which explain high proportion of variability in r_t
- Bai and Ng (Econometrica, 2002) is popular information criteria
- BN criteria involves:
- $\widehat{\sigma}_i^2(m)$ for $i = 1, .., N$ (estimate of error variance for each asset return)
- $\widehat{\sigma}^2(m) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\sigma}_i^2(m)$ (average over all asset returns)
-

$$BN(m) = \widehat{\sigma}^2(m) + m\widehat{\sigma}^2(m)\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)$$

- Estimate factor model for $m = 1, 2, 3..$
- Choose value which minimized $BN(m)$

Selecting the Number of Factors: Scree Plots

- Scree plot is an informal way of selecting number of factors
- To obtain a scree plot in **R**, the `fviz_eig()` command from the `factoextra` package can be used
- Plot eigenvalues (largest to smallest) against number of components
- Look for "elbow" in scree plot
- Eigenvalues at or beyond elbow are small, indicating that factors beyond this point can be ignored
- Next slide: scree plot for our IBM, HPQ, INTC, JPM and BAC
- Elbow occurs at 2, so go with 2 factors


Scree plot of eigenvalues after pca
Eigenvalues
.5
1
1.5
2
2.5
1
2
3
4
5
Number

Summary

- Static factor model is popular in finance
- Often form of factors known (market model)
- Econometric methods same as for regression
- If unknown, principal components can be used
- PC: information in large number of variables can be extracted into small number of factor
- PCs used as regressors (econometric methods same as for regression)