Factor Models and High Dimensional Forecasting

Francis J. DiTraglia

University of Pennsylvania

Econ 722

Survey Articles on Factor Models

Stock & Watson (2010)

Best general overview of factor models and applications.

Bai & Ng (2008)

Comprehensive review of large-sample results for high-dimensional factor models estimated via PCA.

Stock & Watson (2006)

Handbook chapter on forecasting with many predictors. One section is devoted to dynamic factor models.

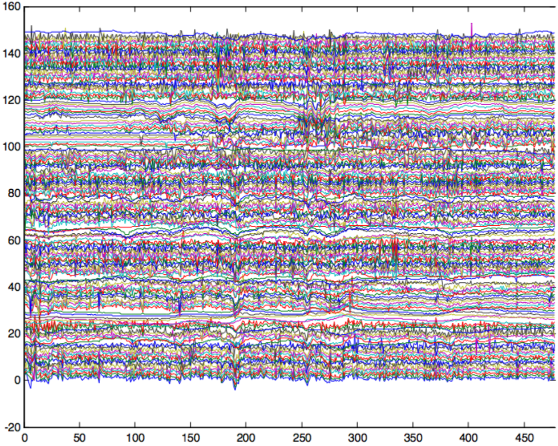
Breitung & Eickmeyer (2006)

Brief overview with an application to Euro-area business cycles.

The Basic Idea

We're interested in settings with a large number of time series N and a comparable number of time periods T.

Example: Stock and Watson Dataset



Monthly Macroeconomic Indicators: $N > 200$, $T > 400$

Why Factor Models?

1. Factors could be intrinsically interesting if they arise from a theoretical model (e.g. Financial Economics)
2. Many variables without running out of degrees of freedom
   - More information could improve forecasts/macro analysis
   - Mimic central banks "looking at everything"
3. Eliminate measurement error and idiosyncratic shocks to provide more reliable information for policy
4. "Remain Agnostic about the Structure of the Economy"
   - Advantages over SVARs: don't have to choose variables to control degrees of freedom, and can allow fewer underlying shocks than variables.

Classical Factor Analysis Model

Assume that X_t has been de-meaned...

$$\underset{(N\times 1)}{X_t} = \Lambda \underset{(r\times 1)}{F_t} + \epsilon_t$$

$$\begin{bmatrix} F_t \\ \epsilon_t \end{bmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_r & 0 \\ 0 & \Psi \end{bmatrix} \right)$$

$\Lambda =$ matrix of factor loadings

$\Psi =$ diagonal matrix of idiosyncratic variances.

Adding Time-Dependence

$$\underset{(N\times 1)}{X_t} = \Lambda \underset{(r\times 1)}{F_t} + \epsilon_t$$

$$\underset{(r\times 1)}{F_t} = A_1 F_{t-1} + \ldots + A_p F_{t-p} + u_t$$

$$\begin{bmatrix} u_t \\ \epsilon_t \end{bmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_r & 0 \\ 0 & \Psi \end{bmatrix} \right)$$

Terminology

Static X_t depends only on F_t

Dynamic X_t depends on lags of F_t as well

Exact Ψ is diagonal and ϵ_t independent over time

Approximate Some cross-sectional & temporal dependence in ϵ_t

The model I wrote down on the previous slide is sometimes called an "exact, static factor model" even though F_t has dynamics.

Some Caveats

1. The difference between "static" and "dynamic" is unclear
   - Can write dynamic model as a static one with more factors
   - Static representation involves "different" factors, but we may not care: are the factors "real" or just a data summary?
2. Not really possible to allow cross-sectional dependence in ϵ_t
   - Unless the off-diagonal elements of Ψ are close to zero we can't tell them apart from the common factors
   - "Approximate" factor models basically assume conditions under which the off-diagonal elements of Ψ are negligible
   - Similarly, time series dependence in ϵ_t can't be very strong (stationary ARMA is ok)

Methods of Estimation for Dynamic Factor Models

1. Bayesian Estimation
2. Maximum Likelihood: EM-Algorithm + Kalman Filter
   - Watson & Engle (1983)
   - Ghahramani & Hinton (1996)
   - Jungbacker & Koopman (2008)
   - Doz, Giannone & Reichlin (2012)
3. "Nonparametric" Estimation
   - Just carry out PCA on X and ignore the time-series element
   - The first r PCs are our estimates $\widehat{F}_t$
   - Essentially treats F_t as an r-dimensional *parameter* to be estimated from an N-dimensional observation X_t

Estimation by PCA

PCA Normalization

- $F'F/T = I_r$ where $F = (F_1, \ldots, F_T)'$
- $\Lambda'\Lambda = \text{diag}(\mu_1, \ldots, \mu_r)$ where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r$

Assumption I

Factors are *pervasive*: $\Lambda'\Lambda/N \to D_\Lambda$ an $(r \times r)$ full rank matrix.

Assumption II

$\text{max e-value } E[\epsilon_t \epsilon_t'] \leq c \leq \infty$ for all N.

Upshot of the Assumptions

If we average over the cross-section, the contribution from the factors persists and the contribution from the idiosyncratic terms disappears as $N \to \infty$.

Key Result for PCA Estimation

Under the assumptions on the previous slide and some other technical conditions, the first r PCs of X consistently estimate the space spanned by the factors as $N, T \to \infty$.

Doz, Giannone & Reichlin (2012)

The arguments for the PCA approach...

- Consistent estimation of factors under very weak assumptions
- MLE is computationally infeasible for large N

...may be somewhat exaggerated.

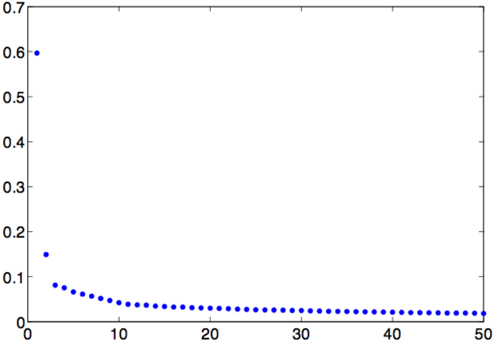
- EM-algorithm + Kalman Filter is *very efficient* – complexity depends on number of *factors*, not number of series
- Treat exact, static factor model (the one I wrote out) as a mis-specified *approximating model* (Quasi-MLE)
- Identical large-sample results as PC under similar assumptions, but better finite-sample properties and temporal smoothing

Choosing the Number of Factors

If we use Likelihood-based or Bayesian estimation, we could try to resort to the familiar tools from earlier in the semester.There are a lot of parameters in factor models, however, so the asymptotic approximations (I'm looking at you, AIC) could be poor.

Choosing the Number of Factors – Scree Plot

If we use PC estimation, we can look a something called a "scree plot" to help us decide how many PCs to include:



This figure depicts the eigenvalues for an $N = 1148$, $T = 252$ dataset of excess stock returns

Choosing the Number of Factors – Bai & Ng (2002)

Choose r to minimize an information criterion:

$$IC(r) = \log V_r(\widehat{\Lambda}, \widehat{F}) + r \cdot g(N, T)$$

where

$$V_r(\Lambda, F) = \frac{1}{NT} \sum_{t=1}^{T} (X_t - \Lambda F_t)'(X_t - \Lambda F_t)$$

and g is a penalty function. The paper provides conditions on the penalty function that guarantee consistent estimation of the true number of factors.

What Can We Do with Factors?

Among other possibilities:

1. Use them to construct Forecasts
2. Use them as Instrumental Variables
3. Use them to "Augment" a VAR

We may not have time for the last two items this semester, so I've put them in the appendix to the slides below...

Some Special Problems in High-dimensional Forecasting

Estimation Uncertainty

We've already seen that OLS can perform very badly if the number of regressors is large relative to sample size.

Best Subsets Infeasible

With more than 30 or so regressors, we can't check all subsets of predictors making classical model selection problematic.

Noise Accumulation

Large N is supposed to help in factor models: averaging over the cross-section gives a consistent estimator of factor space. This can fail in practice, however, since it relies on the assumption that the factors are *pervasive*. See Boivin & Ng (2006).

Main References

Stock & Watson (2006) – "Forecasting with Many Predictors"

Overview of high-dimesional forecasting with a review of forecast combination, factor models, and Bayesian approaches.

Ng (2013) – "Variable Selection in Predictive Regressions"

Reviews and relates a number of shrinkage & selection methods.

Stock & Watson (2012)

Examines a wide range of shrinkage procedures to see if they can improve on diffusion index forecasts.

Kim & Nelson (2013)

"Horse Race" of various factor and shrinkage methods for forecasting.

Diffusion Index Forecasting – Stock & Watson (2002a,b)

JASA paper has the theory, JBES paper has macro forecasting example.

Basic Setup

Forecast scalar time series y_{t+1} using N-dimensional collection of time series X_t where we observe periods $t = 1, \ldots, T$.

Assumption

Static representation of Dynamic Factor Model:

$$\begin{aligned} y_t &= \beta' F_t + \gamma(L) y_t + \epsilon_{t+1} \\ X_t &= \Lambda F_t + e_t \end{aligned}$$

"Direct" Multistep Ahead Forecasts

"Iterated" forecast would be linear in F_t, y_t and lags:

$$y^h_{t+h} = \alpha_h + \beta_h(L) F_t + \gamma_h(L) y_t + \epsilon^h_{t+h}$$

This is really just PCR

Diffusion Index Forecasting – Stock & Watson (2002a,b)

Estimation Procedure

1. Data Pre-processing
   1.1 Transform all series to stationarity (logs or first difference)
   1.2 Center and standardize all series
   1.3 Remove outliers (ten times IQR from median)
   1.4 Optionally augment X_t with lags
2. Estimate the Factors
   - No missing observations: PCA on X_t to estimate $\widehat{F}_t$
   - Missing observations/Mixed-frequency: EM-algorithm
3. Fit the Forecasting Regression
   - Regress y_t on a constant and lags of $\widehat{F}_t$ and y_t to estimate the parameters of the "Direct" multistep forecasting regression.

Diffusion Index Forecasting – Stock & Watson (2002b)

Recall from above that, under certain assumptions, PCA consistently estimates the space spanned by the factors. Broadly similar assumptions are at work here.

Main Theoretical Result

Moment restrictions on (ϵ, e, F) plus a "rank condition" on Λ imply that the MSE of the procedure on the previous slide converges to that of the infeasible optimal procedure, provided that $N, T \to \infty$.

Diffusion Index Forecasting – Stock & Watson (2002a)

Forecasting Experiment

- Simulated real-time forecasting of eight monthly macro variables from 1959:1 to 1998:12
- Forecasting Horizons: 6, 12, and 24 months
- "Training Period" 1959:1 through 1970:1
- Predict h-steps ahead out-of-sample, roll and re-estimate.
- BIC to select lags and # of Factors in forecasting regression
- Compare Diffusion Index Forecasts to Benchmark
  - AR only
  - Factors only
  - AR + Factors

Diffusion Index Forecasting – Stock & Watson (2002a)

Empirical Results

- Factors provide a substantial improvement over benchmark forecasts in terms of MSPE
- Six factors explain 39% of the variance in the 215 series; twelve explain 53%
- Using all 215 series tends to work better than restricting to balanced panel of 149 (PCA estimation)
- Augmenting X_t with lags isn't helpful

What about Ridge and Lasso?

Basic Idea

Diffusion index forecasts are really just PCR. Why not try Ridge or Lasso with all predictors rather than estimating factors?

De Mol, Giannone & Reichlin (2008)

- Compare PCA-based factor forecasts to Ridge and Lasso
- In a small out-of-sample experiment, Ridge and Lasso with appropriate penalty parameters give results comparable to diffusion index.
- Analyze asymptotics of Ridge under assumptions typically used to justify PCA

Other Ways of Extracting Factors

Sparse PCA

Add a Lasso-type penalty to the "regression" formulation of PCA: encourage the factors to load on small number of variables.

Independent Components Analysis (ICA)

Extract factors that maximize non-Gaussianity

Both of these are considered in Kim & Swanson (2014) and seem to work very well when combined with second-stage shrinkage.

To Target or Not to Target?

Problem with PCA and Friends

Completely ignores Y in constructing the factors! Should we take the forecast target into account when extracting factors?

Some References

- Bai & Ng (2008) – Forecasting Economic Time Series Using Targeted Predictors
- Kelly & Pruitt (2012) – The Three-pass Regression Filter

Partial Least Squares (PLS)

As an Optimization Problem

Construct a sequence of linear combinations of X that solve

$$\max_{\alpha} Corr^2(\mathbf{y}, X\alpha) Var(X\alpha)$$

subject to $||\alpha|| = 1$ and the constraint that each PLS "factor" is orthogonal to the preceding ones.

As a Probabilistic Model

"Shared" factor F_t and X-specific factor Z_t

$$\begin{aligned} Y_t &= \mu_Y + \Lambda_Y F_t + \epsilon_t \\ X_t &= \mu_X + \Lambda_X F_t + \Pi Z_t + u_t \end{aligned}$$

where $F_t \perp Z_t$

Bootstrap Aggregation – "Bagging"

Bagging Algorithm

1. Make a bootstrap draw
2. Carry out selection/shrinkage/estimation using boostrap data
3. Use estimated parameters from to construct a forecast $\widehat{y}_{T+h}^{(b)}$
4. Repeat for $b = 1, \ldots, B$
5. Average to get "Bagged" Forecast: $\widehat{y}_{T+h}^{(Bag)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{y}_{T+h}^{(b)}$

Details

- If the data are dependent, need block bootstrap.
- In step 3, we forecast using the *parameters* estimated from the bootstrap data but the *predictors* from the *real* dataset.

Bootstrap Aggregation – “Bagging”

Why Bagging?

- Aims to reduce the forecast error of “unstable” procedures such as variable selection of Lasso, by reducing their variance.
- Completely portable: you can bag *anything* provided you have an appropriate way to carry out the bootstrap.
- May provide a way of attacking the problem of inference post-model selection. See Efron (JASA, *Forthcoming*) “Estimation and Accuracy after Model Selection”

Bagging in Economics

Inoue & Killian (2008, JASA)

Compares performance of bagged "pre-test" estimator (variable selection via a t-test) to other methods of forecasting US Inflation. Bagging is carried out via a block bootstrap.

Stock & Watson (2012)

Among other shrinkage procedures, they consider a large-sample approximation to bagging pre-test estimators that doesn't require making bootstrap draws.

Other Papers That Use Bagging

- Hillebrand & Medeiros (2010): Realized Volatility Forecasts
- Hillebrand et al (2012): Forecasting the Equity Premium
- Kim and Swanson (2013)

Boosting

Ensemble Methods

Machine learning term for "non-Bayesian model averaging"

What is Boosting?

- Combine large number of "weak learners" (i.e. crappy predictive models) so that the *ensemble* predicts well.
- Explicitly designed around predictive loss
- Arbitrarily improve in-sample fit of arbitrarily the weak learners!

Book-Length Treatment

Shapire & Freund (2012) – *Boosting: Foundations and Algorithms*

Boosting

Bai & Ng (2009) – Boosting Diffusion Indices

Use boosting to select which lags of factors to include in a forecasting regression estimated following PCA.

Buchen & Wohlrabe (2011) – Is Boosting a Viable Alternative?

Boosting performs well compared to other methods in the example from the 2006 Stock & Watson Handbook Chapter.

Ng (2014) – Boosting Recessions

Appendix

Factors as Instruments – Bai & Ng (2010)

Endogenous Regressors x_t

$$y_t = x_t'\beta + \epsilon_t \qquad E[x_t\epsilon_t] \neq 0$$

Unobserved Variables F_t are Strong IVs

$$\underset{(k\times1)}{x_t} = \Psi' \underset{(r\times1)}{F_t} + u_t \qquad E[F_t\epsilon_t] = 0$$

Observe Large Panel $(z_{1t}, \ldots, z_{Nt})$

$$z_{it} = \lambda_i' F_t + e_{it}$$

Factors as Instruments – Bai & Ng (2010)

$$\boxed{y_t = x_t'\beta + \epsilon_t, \qquad x_t = \Psi' F_t + u_t, \qquad z_{it} = \lambda_i' F_t + e_{it}}$$

Procedure

1. Calculate the PCs of Z
2. Calculate $\widetilde{F}_t$ using the first r PCs of Z
3. Use $\widetilde{F}_t$ in place of F_t for IV estimation

Main Result

Under certain assumptions, as $(N, T) \to \infty$ "estimation and inference can proceed as though F_t were known." The resulting estimator is consistent and asymptotically normal.

Factors as Instruments – Bai & Ng (2010)

Why Might This be Helpful?

1. Avoid many instruments bias
2. Avoid bias from irrelevant instruments
3. Allow more observed instruments z_{it} than sample size T
4. Provided that $\sqrt{T}/N \to 0$, all of the observed instruments z_{it} can be *endogenous* as long as F_t is exogenous

FAVARs – Bernanke, Boivin & Eliasz (2005)

Two Problems with Structural VARs

1. Number of parameters is *quadratic* in the number of variables. Unrestricted VAR infeasible unless T is large relative to N.
   - You've studied one solution to this problem already this semester: Bayesian Estimation with informative priors
2. To keep estimation tractable we typically use a small number of variables, but then the VAR innovations "might not span the space of structural shocks."

FAVARs – Bernanke, Boivin & Eliasz (2005)

Factor-Augmented VAR Model

$$\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \Phi(L) \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + v_t$$

$$X_t = \Lambda^f F_t + \Lambda^y Y_t + e_t$$

$\underset{(M \times 1)}{Y_t}$ = observable variables that "drive dynamics of the economy"

$\underset{(K \times 1)}{F_t}$ = Small # of unobserved factors: "additional information"

$\underset{(N \times 1)}{X_t}$ = Large # of observed "informational time series"

FAVARs – Bernanke, Boivin & Eliasz (2005)

$$\boxed{\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \Phi(L) \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + v_t \qquad X_t = \Lambda^f F_t + \Lambda^y Y_t + e_t}$$

Consider Two Estimation Procedures

1. Two-step Procedure:
   - Estimate space spanned by factors using first $K + M$ PCs of X
   - Estimate VAR with $\widehat{F}_t$ in place of F_t
2. Full Bayes (Gibbs Sampler)

Empirical Application

Additional information contained in FVAR is "important to properly identify the monetary transmission mechanism."